Contents

Original Papers

An Algorithm to Classify Real-World Ambulatory Status From a Wearable Device Using Multimodal and Demographically Diverse Data: Validation Study (e43726)
Sara Popham, Maximilien Burq, Erin Rainaldi, Sooyoon Shin, Jessilyn Dunn, Ritu Kapur ................................................................. 2

The Variability of Lumbar Sequential Motion Patterns: Observational Study (e41906)
Inge Caelers, Toon Boselie, Wouter van Hemert, Kim Rijkers, Rob De Bie, Henk van Santbrink ............................................................... 16

Development and Testing of a Data Capture Device for Use With Clinical Incentive Spirometers: Testing and Usability Study (e46653)
Michael Burns, Anik Sinha, Alexander Hoffmann, Zewen Wu, Tomas Medina Inchauste, Aaron Retsky, David Chesney, Sachin Kheterpal, Nirav Shah .................................................................................................................. 46

Review

Mixed Reality Platforms in Telehealth Delivery: Scoping Review (e42709)
Hemendra Worlikar, Sean Coleman, Jack Kelly, Sadhbh O’Connor, Aoife Murray, Terri McVeigh, Jennifer Doran, Ian McCabe, Derek O’Keeffe. 2 5
An Algorithm to Classify Real-World Ambulatory Status From a Wearable Device Using Multimodal and Demographically Diverse Data: Validation Study

Sara Popham 1, PhD; Maximilien Burq 1, PhD; Erin E Rainaldi 1, MSc; Sooyoon Shin 1, PhD; Jessilyn Dunn 2,3,4, PhD; Ritu Kapur 1, PhD

1Verily Life Sciences, South San Francisco, CA, United States
2Department of Biomedical Engineering, Duke University, Durham, NC, United States
3Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, United States
4Duke Clinical Research Institute, Durham, NC, United States

Corresponding Author:
Sara Popham, PhD
Verily Life Sciences
269 E Grand Ave
South San Francisco, CA, 94080
United States
Phone: 1 650 253 0000
Email: spopham@verily.com

Abstract

Background: Measuring the amount of physical activity and its patterns using wearable sensor technology in real-world settings can provide critical insights into health status.

Objective: This study’s aim was to develop and evaluate the analytical validity and transdemographic generalizability of an algorithm that classifies binary ambulatory status (yes or no) on the accelerometer signal from wrist-worn biometric monitoring technology.

Methods: Biometric monitoring technology algorithm validation traditionally relies on large numbers of self-reported labels or on periods of high-resolution monitoring with reference devices. We used both methods on data collected from 2 distinct studies for algorithm training and testing, one with precise ground-truth labels from a reference device (n=75) and the second with participant-reported ground-truth labels from a more diverse, larger sample (n=1691); in total, we collected data from 16.7 million 10-second epochs. We trained a neural network on a combined data set and measured performance in multiple held-out testing data sets, overall and in demographically stratified subgroups.

Results: The algorithm was accurate at classifying ambulatory status in 10-second epochs (area under the curve 0.938; 95% CI 0.921-0.958) and on daily aggregate metrics (daily mean absolute percentage error 18%; 95% CI 15%-20%) without significant performance differences across subgroups.

Conclusions: Our algorithm can accurately classify ambulatory status with a wrist-worn device in real-world settings with generalizability across demographic subgroups. The validated algorithm can effectively quantify users’ walking activity and help researchers gain insights on users’ health status.

(JMIR Biomed Eng 2023;8:e43726) doi:10.2196/43726

KEYWORDS
digital measurement; wearable sensor; machine learning; ambulatory status; Project Baseline Health Study; physical activity

Introduction

Quantifying physical activity can be highly informative about both general health status and the condition of people with specific diseases [1,2]. Characteristics of physical activity have been shown to be prognostic factors in various chronic conditions [3-13]. Yet reliably producing research-grade measurements of physical activity in real-world settings remains a challenge. Traditionally, the validation of such measurements often relies on individual self-reports or is performed...
Recently, the advent of wearable technology has made it possible to measure physical activity to a previously untenable extent [17,18]. Ambulatory activity in particular, namely whether individuals are walking and how much, is a basic aspect of physical activity that can be investigated in general populations and in specific clinical settings. Wearable devices can collect information passively during daily living and generate a vast quantity of digital measurements that allow researchers to probe functional physical activity generally and ambulatory activity specifically. Using these digital measures in research studies, however, requires analytical validation [19]. In their design, validation studies have to balance factors such as feasibility and the resource-intensiveness of their data collection approach with demonstrating validity in representative populations.

To date, the majority of measurements in validation studies have come from either short observation periods in laboratory settings [20,21] or self-reported labels in real-world settings [22]. Laboratory measurements often render observations with exceptionally clean and easy-to-use ground-truth labels, but algorithms trained on data of this kind do not always generalize to everyday activities [23]. On the other hand, using self-reported labels as the ground truth yields a closer reflection of individual everyday activities, but these labels are often noisy and less accurate [15,16,24]. There have been some examples of reference devices deployed to generate accurate truth labels in generalizable real-world settings [25,26], but this came at the cost of intrusiveness and resource-intensive data processing steps after collection, such as manual video footage tagging. With all these considerations in mind, validation studies tend to be highly heterogeneous, and need to be interpreted in context.

Herein we report on the development and analytical validation of an ambulatory status classification algorithm. This algorithm classifies the ambulatory status of users of a wrist-worn device in real-world environments. We carried out 2 separate studies including participants from independent populations with distinct sources of ground-truth labels for a deeper characterization of the algorithm performance. One of the studies, the pilot program study, used a relatively small and demographically homogeneous cohort, where participants provided a highly accurate ground-truth source from a reference device. The other study was derived from the Project Baseline Health Study (PBHS), a prospective, multicenter, longitudinal study with participants of diverse backgrounds who were representative of the entire health spectrum [27]; this was a demographically diverse cohort that provided self-reported labels as the ground-truth source. This cohort was also relatively large, and we therefore expected it to yield results less susceptible to outlier readouts. We present analytical validation results of the performance of our algorithm against the highly accurate ground-truth source (from the pilot), and we examine the generalizability of the results across a study population of demographically diverse individuals (in the PBHS).

**Methods**

### Participant Cohorts

Two distinct studies were conducted, with training and testing groups identified a priori within each study. Participants in both studies wore the smartwatch (the Verily Study Watch) [27-30].

The first study was a pilot program (n=75) of adult volunteer participants recruited among Verily Life Sciences employees in 2 locations (South San Francisco, California, and Cambridge, Massachusetts) without specific selection criteria. For this group, ground-truth labels were collected from an ankle-worn reference device (StepWatch 4). The Verily Study Watch and reference device were worn simultaneously for 7 consecutive days to ensure capture of both weekday and weekend behavior; for each participant, days were included as evaluable if both devices were worn synchronously for a minimum of 8 hours. No demographic information on race or ethnicity was collected in this study. The observation period ran from June to December 2019.

In order to expand the demographic representativeness of the overall validation effort, the second study included a large and diverse cohort (n=1691) consisting of participants from the PBHS who consented to participate in this substudy [27]. The period for data collection ran from May to December 2019.

### Ethics Approval

The pilot program was determined to be exempt research that did not require institutional review board review. Written informed consent was obtained from all participants enrolled in the PBHS; the PBHS was approved by both the WCG institutional review board (approval tracking number 20170163, work order number 1-1506365-1) and the institutional review boards of each participating institution (Stanford University, Duke University, and the California Health and Longevity Institute) [27]. The PBHS was registered at ClinicalTrials.gov (NCT03154346).

All methods complied with relevant guidelines and regulations; the research involving human participants was performed in accordance with relevant guidelines and regulations. Experimental protocols were approved by appropriate committees from Verily Life Sciences and by PBHS governance (participating institutions are above).

### Wearable Devices

The Verily Study Watch recorded acceleration data in both cohorts via an onboard inertial measurement unit with a 30 Hz 3-axis accelerometer. For the PBHS population, the smartwatch also contained a user interface allowing participants to tag their activities on the watch (Figure 1A).
The reference device for the pilot program was an ankle-worn single-axis accelerometer (Modus StepWatch 4) that provided step count as a reference label for algorithm development.

**Reference Labels**

In the pilot program, we generated reference labels on data collected from the ankle-worn StepWatch: 10-second windows were considered “ambulatory” if they had ≥3 steps on the wearing foot and “nonambulatory” if they had <3 steps [31]. The default window size returned by this device was 10 seconds, and this was deemed to provide good temporal granularity.

For the self-reported reference labels from the PBHS, participants tagged their activities as 1 of 3 options listed by the wrist device: “walk/run,” “still,” and “other.” Participants could tag the start and end of an activity period directly on the study device.
watch, which enabled precise synchronization of the labels to the raw sensor data stream. When necessary, participants could edit or delete tags as needed (Figure 1B). For the purpose of this analysis, “still” and “other” were grouped together under the “nonambulatory” label, while “walk/run” was equated to “ambulatory.”

The amount of data used from each of these studies is summarized in Multimedia Appendix 1, Table S3.

Algorithm Development

Data from each study (pilot program, n=75; PBHS, n=1691) were split into nonoverlapping training and testing data sets at the participant level. For each study, data from approximately half the participants were used for training the algorithm and data from the other half were held out for algorithm testing. We decided on a 50-50 split in order to retain statistical power in the testing data, particularly considering the intended additional analyses of different demographic subgroups (discussed below).

In the pilot program, the split into training and testing data sets was based on participants’ daily step counts in order to mitigate potential algorithmic biases caused by training primarily on data from participants with either very low or very high activity levels. The difference in the mean daily step counts between the 2 halves of the split was 234 steps. For the PBHS cohort, the split into training and testing data sets was done randomly, as participants did not have daily aggregated results. We trained multiple versions of the algorithm with combinations of different subsets of training data and compared performance across these different algorithms (Figure 1).

We developed an algorithm that classifies the ambulatory status of device users in 10-second epochs (as ambulatory vs nonambulatory). First, the following 14 features were extracted from the Verily Study Watch’s acceleration data, in 10-second epochs: 3 features related to deviations of the signal, 5 features derived from power spectral density energy in frequency bands typically associated with user ambulation (ie, walking or running), 2 features that are signal percentiles (ie, 95th percentiles), and 4 features that are differences between signal percentiles (ie, IQR). These features were fed into a shallow neural network model with 2 dense layers: ReLu nonlinearities and softmax of outputs. The neural network was trained with a batch size of 32. The Adam optimizer was used with a learning rate of 0.001, and loss was calculated using categorical cross-entropy. Training ran for 10 epochs. Alternative features and neural network architectures were explored using the training data, but larger feature sets or more complex architectures did not result in higher performance, so this algorithm was chosen.

The classifier threshold was optimized to minimize absolute percentage error on daily ambulatory time on the training data from the pilot study (vs the data from the reference device used as the ground-truth source, as discussed above). For this optimization process, we performed 5-fold cross-validation at the participant level within the training data. We found the minimum daily mean absolute percentage error (MAPE) across the aggregated held-out data from all folds using a 1D grid search procedure.

The signal-processing, feature selection, model training, and hyperparameter tuning were all performed on training data sets identified a priori.

Analyses

The demographic characteristics of the study cohorts were analyzed using descriptive statistics.

We analyzed the following metrics to characterize the performance of the algorithm, calculated on the held-out test sets: area under the receiver operating characteristic curve (AUC) for the overall study cohorts and across different demographic subcohorts within the PBHS cohort (this was chosen as the metric for comparison because, unlike other measures, such as F1-score or accuracy, it is not susceptible to differences in the chosen classifier threshold), mean accuracy, and MAPE of daily ambulatory time, defined as the summing of all 10-second windows that were labeled as “ambulatory” in a day.

Analyses were performed in python using NumPy (version 1.21.5), pandas (version 1.1.5), SciPy (version 1.2.1), scikit-learn (version 1.0.2), and tensorflow (version 2.10.0).

Confidence intervals were calculated using the bootstrap method with 1000 resampling iterations. Resampling was done at the participant level to ensure that all data from a single participant were either included or excluded within each resampling iteration.

Results

Characteristics of Participants From the Pilot Study and the PBHS Cohort

Participants in the pilot study were mostly male (45/75, 64%), with a mean age of 33 (SD 8.5) years. Participants from the PBHS were more often female (1366/2502, 55%), with a mean age of 54 (SD 17) years (Multimedia Appendix 1, Table S1).

Algorithm Training

Data from each study were separately split (approximately 50-50) into nonoverlapping training and testing data sets (Figure 1); this allocation was done at the participant level (n=75 from the pilot study and n=1691 from the PBHS population). Out of 16,769 million 10-second epochs collected from the 2 studies, 8,841 million 10-second epochs were used for training across all algorithm iterations generated (the data sets are described in Multimedia Appendix 1, Table S3).

From the pilot program study, a total of 1,641,272 nonoverlapping 10-second epochs were collected (n=70 participants; Figure 1), of which 228,721 (13.9%) were “ambulatory” according to the reference device–based labels. We used 879,593 10-second epochs (from 35 unique participants) for training (118,730, 13.5% of which were “ambulatory”; Multimedia Appendix 1, Table S3).

We collected a total of 14,814,910 nonoverlapping 10-second epochs from the PBHS (n=1531 participants; Figure 1), of which 7,079,216 (47.8%) were “ambulatory” according to the participant-reported reference labels. The proportion of
“ambulatory” labels in the PBHS was higher than in the pilot program study (47.8% vs 13.9%), which is likely attributable to the different labeling methods across studies. We expect that labeling from the pilot study was more stringent to show true ambulatory epochs, because these were determined directly by the reference device readouts (i.e., any 10-second epoch with greater than or equal to 6 steps, relative to all 10-second epochs collected during the wear time). In the PBHS, the proportion of ambulatory labels was determined based on participant self-reported, manually entered walk/run tags relative to all entered tags. PBHS tagging, therefore, can be more vulnerable to selection bias toward “ambulatory,” since participants may favor reporting active over inactive states.

Data from the PBHS were not only divided into training and testing sets, but, across each set, we considered 2 quality control (QC) strata to test the impact of data quality on the development and performance of the algorithm. An extremely light QC selection, eliminating labels with gross apparent user errors (such as tags that were longer than a full day), was applied to generate the “QC-minimal” sub-data set, which therefore included virtually all labels suitable for evaluation (10,264/104,212, 9.8% of user-tagged events were eliminated, and another 12,010/104,212, 11.5% were truncated); a more stringent selection was applied to generate the “QC-high” sub-data set (80,852/104,212, 77.6% of user-tagged events were eliminated, and all tags were truncated to some degree; Figure 1 and Multimedia Appendix 1, Table S2). The 2 strata aimed to parse out performance variability due to noise generated by imperfectly self-reported reference labels (this was not a factor for the labels from the reference device in the pilot program).

The resulting size of these QC training sub-data sets was 160,778 10-second epochs for QC-high (n=173 participants) and 7,802,829 for QC-minimal (n=829 participants). Of these labeled epochs, 102,783 (63.7%) and 3,863,964 (49.5%), respectively, were ambulatory according to the participant-reported tags (Figure 1 and Multimedia Appendix 1, Table S3).

**Effect of Raw Data Quality on Algorithm Performance**

We tested each of the algorithm iterations from the training process above (originated using the 2 PBHS QC-sub-data sets and the pilot data set) across data from the held-out QC sub-data sets from the PBHS and the pilot program by calculating AUC values across all combinations. Namely, we tested each of the following algorithms against the held-out data sets from the pilot study and the PBHS QC-high and QC-minimal sub-data sets (Figure 2): (1) trained with the PBHS QC-high sub-data set, (2) trained with the PBHS QC-minimal sub-data set, (3) trained with the pooled PBHS QC-high plus pilot data set, (4) trained with the pooled PBHS QC-minimal plus pilot data set, and (5) trained with just the pilot data set. For each algorithm iteration, AUC values varied across the testing sub-data sets (QC sub-data sets from the PBHS and pilot program), with differences ranging from 0.047 to 0.187. For each test data set, the AUC variations across the algorithm iterations (1) through (5) were narrower, with differences ranging between 0.001 and 0.045. Therefore, data quality differences across the training sub-data sets did not appear to affect algorithm performance, as reflected in AUC variability, as much as data quality in the testing sub-data sets.

**Figure 2.** (A) Heat map of AUC values for the algorithm iterations generated via different training sub-data sets from the PBHS when tested on each of the separate testing cohorts. (B) AUC values for the algorithm iterations generated via different training sub-data sets from the PBHS when tested on each of the separate testing cohorts, with error bars based on the 95% CI. Each testing cohort is shown with a different color or symbol. From top to bottom, the red dotted lines indicate mean AUC values for the pilot, PBHS QC-high, and PBHS QC-minimal test data sets, respectively. The model trained on combined PBHS QC-high and pilot training data (highlighted in yellow) was the version of the algorithm used for further analyses. AUC: area under the receiver operating characteristic curve; PBHS: Project Baseline Health Study; QC: quality control.

Based on the testing results described above, we selected an algorithm trained using combined data from one of the PBHS sub-data sets (QC-high) plus the pilot program data set to proceed to further analysis. This algorithm iteration (termed “version 2022”) showed the highest testing performance (evaluated by AUC) calculated with data from the pilot program (the most precise and cleanest data set) without substantially reduced performance on PBHS data (Figure 2). With this approach, we prioritized testing the accuracy of the algorithm against participants’ actual ambulatory status based on the
reference device, not against the type of labels that are most feasible to obtain (ie, self-reported labels), although we report accuracy on both types of labels.

**Algorithm Testing**

Tested against the held-out data set from the pilot program (Table 1), the selected algorithm had a sensitivity of 71% and a specificity of 95%, for an overall accuracy of 91.5% (95% CI 90.3%-92.9%; Figure 3A) and an AUC of 0.938 (95% CI 0.921-0.958; Figure 3B) when classifying the ambulatory status of 10-second epochs. When tested on the held-out data set from the PBHS QC-high sub-data set, the selected algorithm had an overall accuracy of 75.7% (95% CI 72.5%-78.6%) and an AUC of 0.832 (95% CI 0.800-0.864).

**Table 1.** Algorithm performance measures.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV&lt;sup&gt;a&lt;/sup&gt;</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;-score</th>
<th>AUC-ROC&lt;sup&gt;b&lt;/sup&gt;</th>
<th>AUC-PRC&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot study</td>
<td>91.3%</td>
<td>0.706</td>
<td>0.948</td>
<td>0.696</td>
<td>0.701</td>
<td>0.938</td>
<td>0.781</td>
</tr>
<tr>
<td>PBHS QC&lt;sup&gt;d&lt;/sup&gt;-high</td>
<td>75.8%</td>
<td>0.731</td>
<td>0.802</td>
<td>0.885</td>
<td>0.788</td>
<td>0.832</td>
<td>0.901</td>
</tr>
</tbody>
</table>

<sup>a</sup>PPV: positive predictive value.

<sup>b</sup>AUC-ROC: area under the receiver operating characteristic curve.

<sup>c</sup>AUC-PRC: area under the precision-recall curve.

<sup>d</sup>PBHS: Project Baseline Health Study.

<sup>e</sup>QC: quality control.

**Figure 3.** (A) Accuracy of the algorithm selected for full analysis, as evaluated in the pilot cohort. Here, the color map denotes K, the number of 10-second epochs. Percentages are normalized across rows, which allows easy reading of the sensitivity and specificity values. (B) Receiver operating characteristic curve and area under the curve of the algorithm selected for full validation, as evaluated in the pilot cohort. The red X denotes the true positive rate and false positive rate of the algorithm at the chosen classifier threshold. AUC: area under the receiver operating characteristic curve.

The proportion of predicted ambulatory epochs of the selected algorithm varied with the number of steps in the 10-second epochs (Figure 4). The lowest proportion of predicted ambulatory epochs happened in the 3 to 5 step range (36%-57% sensitivity, ie, correct predictions as “ambulatory: yes”), and the proportion of epochs classified as ambulatory grew with additional steps in the 10-second window (67%-91% correct predictions). Note that data from epochs with more than 11 recorded single-leg steps are not shown due to their low frequency (the number of samples per step count is shown in Multimedia Appendix 1, Figure S1).
Figure 4. Predictions of the selected algorithm to classify 10-second epochs as ambulatory (or not) according to the number of steps in the 10-second epochs based on the reference device data from the pilot program study. A perfectly performing algorithm would predict “ambulatory” for all epochs with 3 or more steps on the wearing foot (indicated by the blue shadow), and nonambulatory for all epochs with fewer steps (indicated by the gray shadow). Epochs with more than 11 recorded steps are not shown due to their small sample size (Multimedia Appendix 1, Figure S1).

When considering daily step aggregates as the metric of interest, there was good agreement between the algorithm classifications and the reference ($R^2=0.771$), with a MAPE in daily ambulatory time (minutes) of 18% (95% CI 15%-20%) and a median absolute percentage error of 14% (Figure 5A and Figure 5B). The mean absolute error (MAE) of daily ambulatory time was 19.5 (95% CI 15.0-23.2) minutes, and the median absolute error was 14 minutes (Figure 5C). Consistent with the observations at the 10-second epoch level, the magnitude of error in daily ambulatory time (ie, the difference between algorithm-predicted and actual values) was dependent on the actual daily ambulatory time (as computed by the StepWatch; Figure 5D): the chance for underestimating daily ambulatory time (in minutes) grew as the reference daily ambulatory time increased. The largest underestimation we observed was 138.5 minutes in absolute time (relative error 32.5%).
Performance of the Ambulatory Status Classification Algorithm Across Demographic Subgroups

In order to characterize the generalizability of the algorithm’s performance, we calculated AUC values for the selected algorithm across demographic subgroups based on gender, age, and race. Initially, in the testing data set from the pilot program (Figure 6A), the results suggested a possible difference in performance between male and female participants, as seen in the lack of overlap of the 95% CIs. However, in a similar analysis using the larger and more diverse testing data set from the PBHS, which enabled subanalyses by participant gender, race, and age, that difference was no longer present and the results showed no meaningful performance difference across any of the subgroups of age, gender, or race, as evidenced by the overlapping 95% CIs (Figure 6B). A replication of the majority population from the pilot study within the PBHS showed an AUC of 0.8166 (95% CI 0.7501-0.8666) for White males aged 31 to 65 years in the PBHS cohort, which was not significantly different from the AUC of the PBHS cohort as a whole (AUC 0.8339).
Figure 6. Performance (AUC values) of the selected algorithm across different demographic subgroups. (A) The pilot study testing data set. (B) The PBHS QC-high testing data set. AUC: area under the receiver operating characteristic curve; PBHS: Project Baseline Health Study; QC: quality control.

Discussion

This study presents the analytical validation in a real-world setting of an algorithm to classify the ambulatory status of users wearing a smartwatch. The algorithm performs well, distinguishing between ambulatory and nonambulatory states with high accuracy (75.7%-91.5% depending on the testing data set). Furthermore, the approach taken to analytic validation allowed us to investigate multiple subgroups, including age, gender, and race, demonstrating that the high performance of the algorithm is generalizable across a broad range of demographics.

All existing validation studies of ambulatory status classification from wrist-worn sensors have been either performed on young and healthy populations [25] or in the laboratory or clinic [20,21]. Yet measuring ambulatory status or daily ambulatory time is most clinically relevant for people with walking impairments—whether due to age, movement disorders, cardiovascular illness, or other circumstances—and most informative when done in an individual’s own environment (ie, their real-world setting). Thus, a key innovation in this work is our focus on using data captured in real-world settings (as opposed to highly controlled clinic or laboratory settings) from demographically diverse cohorts for the actual development and validation of this algorithm.

Therefore, the novel contributions of this work are 2-fold. First, we introduce a scalable framework for collecting reference labels on ambulatory status via a reference device and via user-reported data for training and validation. As part of that approach, we used 2 separate and different modalities to measure ground-truth status. This strategy enabled us to handle both comprehensive and highly precise labels (in the pilot program), as well as a larger volume of inherently noisy ones (user-reported tags from the PBHS), both in real-world settings. Our strong results across both sets of data indicate that this innovative multimodal approach contributed to a robust development scheme that may have boosted the performance of the resulting algorithm. The long-term practical convenience of a wrist-worn device (as opposed to an ankle-worn device or a dedicated assessment period) may be advantageous for this type of continuous generalizable monitoring [32-35], although a thorough side-by-side analysis of these 2 reference standard measurement methods to fully understand their correlation remains as a topic for future studies.

Second, we leveraged this framework to provide large-scale validation of the performance of the selected algorithm iteration, addressing shortcomings in terms of generalization previously reported in the literature [20,21,32]. Prior studies have used algorithms to report on differences in physical activity by different demographic subgroups but lacked validation data for those algorithms across demographic subgroups [25,36-38]. To our knowledge, this is one of the first studies to show a proper validation approach to develop and test a generalizable algorithm across demographic subgroups where algorithm output could have differed by subgroup.

In addition, our approach highlights several points of interest when developing validation methodologies for this type of algorithm. The increased sample sizes and variability in data quality achieved by combining 2 distinct data sets enabled deeper characterization of the algorithm’s performance. One of our studies generated data sets where truth labels were of high quality and accuracy but were collected from a study population limited in scope; the other study collected data from a large and demographically diverse cohort (albeit a somewhat engaged and self-selected participant group who volunteered and expressed interest in the PBHS and its health technology aspects), which allowed us to conduct subgroup analyses for both training and testing. Our results reinforce the well-established fact that modern machine-learning algorithms can sometimes perform well even when trained on a noisy data set [39]. This observation may be useful for researchers navigating study design decisions and tradeoffs, including...
sample sizes and data labeling methods. For future research, determining the role of data quality factors in the development and characterization of this type of algorithm is an open issue [18].

Our approach to the generation of reference labels was pragmatic, using deployment-friendly ankle-worn devices or user-reported tags. Neither of these was as resource-demanding as other intensive approaches (ie, video monitoring), but generated information of sufficient quality to conduct our validation with relatively high time resolution (10-second epochs). Of note, the intrinsic nature of the 2 methods used for the generation of reference labels probably contributed to the noticeable difference in the proportions of “ambulatory” labels between the 2 studies (discussed in the Results), with the proportion observed in the pilot program study being the one closest to other literature reports [40].

When interpreting our results in the context of existing literature, it is worth noting that most validation studies for this type of algorithm have used step counts as the metric of interest [31,36,41-53], while ambulatory time (or a related metric) is the focus of a minority of reports [54,55]. In general, considering the close correlation between step count and ambulatory time, the performance of our algorithm could be placed on par with other algorithms, yet detailed side-by-side appraisals of results remain challenging; this research field is in need of standardization [19,56,57].

This study also had limitations. First, in principle, the StepWatch readouts used as ground truth may not have provided perfect accuracy, even though there is extensive literature supporting the use of StepWatch as a reference device [31,50,51,56,58-60]. Second, we observed fluctuations in the ambulatory status classification algorithm performance based on daily ambulatory time; this fluctuation was present when the algorithm detected 10-second epochs as ambulatory (or not) and was also manifested in the daily aggregates of ambulatory time. While this trend (shown in Figure 5) may have been driven, partially, by outlying data points with high step counts in our sample, which would be of little relevance in hypothetical clinical scenarios, it may also have been due to low-step periods containing mixed activities in which walking was not the only or dominant source of hand motion. In addition, while the cutoff used to read the StepWatch ambulatory classification relied on existing literature [61], it may not be perfect in itself. In this regard, it could be reassuring that the algorithm handled epochs with step counts between 4 and 8 as a continuum, as this is possibly reflective of the complexities of organic movement.

In sum, we have developed an accurate algorithm for the detection of the ambulatory status of users of a wrist-worn device in a free-living, real-world setting; the output is generalizable across several user demographic characteristics. The characterization of this algorithm was conducted in 2 distinct data sets, which lends credibility to the robustness and applicability of the performance results obtained in this study and illustrates the advantages of similar approaches to future research in this field.

Acknowledgments
The authors wish to thank David Andresen, Anthony Chan, Chen Chen, Robin Lin, Stephen Lanham, David Miller, Shannon Fong, and the Project Baseline Health Study team and participants for their contributions to this work. The authors also wish to acknowledge writing and editing support from Julia Saiz (Verily). This study was sponsored by Verily Life Sciences, which was responsible for data collection. Authors employed by Verily Life Sciences had access to the raw study data in full. All authors were responsible for interpretation of results and writing and review of the manuscript; all authors approved the final manuscript for submission.

Data Availability
The deidentified PBHS data corresponding to this study are available upon request for the purpose of examining their reproducibility. Interested investigators should direct requests to jsaiz@verily.com. Requests are subject to approval by the Project Baseline Health Study governance. Data from the pilot program are not available due to the nature of this program. Participants in this program did not consent for their data to be shared publicly.

Authors’ Contributions
ER and RK contributed to study concept and design; ER contributed to data collection; and SP, MB, SS, ER, and RK contributed to data analysis and interpretation.

Conflicts of Interest
SP, MB, ER, SS, and RK report employment and equity ownership in Verily Life Sciences. JD is a scientific advisor to Veri, Inc.

Multimedia Appendix 1
Supplementary materials.
[DOCX File, 61 KB - biomedeng_v8i1e43726_app1.docx]
References


The Variability of Lumbar Sequential Motion Patterns: Observational Study

Inge Caelers1,2,3*, MD; Toon Boselie1,3*, MD, PhD; Wouter van Hemert1*, MD, PhD; Kim Rijkers1,3*, MD, PhD; Rob De Bie4*, PhD, PT; Henk van Santbrink1,2,3*, MD, PhD

1Department of Neurosurgery, Zuyderland Medical Center, Heerlen, Netherlands
2Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands
3Department of Neurosurgery, Maastricht University Medical Center +, Maastricht, Netherlands
4Department of Epidemiology, Maastricht University, Maastricht, Netherlands
*all authors contributed equally

Corresponding Author:
Inge Caelers, MD
Department of Neurosurgery, Zuyderland Medical Center
Henri Dunantstraat 5
Heerlen, 6419 PC
Netherlands
Phone: 31 652593545
Email: inge.caelers@mumc.nl

Abstract

Background: Physiological motion of the lumbar spine is a topic of interest for musculoskeletal health care professionals since abnormal motion is believed to be related to lumbar complaints. Many researchers have described ranges of motion for the lumbar spine, but only few have mentioned specific motion patterns of each individual segment during flexion and extension, mostly comprising the sequence of segmental initiation in sagittal rotation. However, an adequate definition of physiological motion is still lacking. For the lower cervical spine, a consistent pattern of segmental contributions in a flexion-extension movement in young healthy individuals was described, resulting in a definition of physiological motion of the cervical spine.

Objective: This study aimed to define the lumbar spines’ physiological motion pattern by determining the sequence of segmental contribution in sagittal rotation of each vertebra during maximum flexion and extension in healthy male participants.

Methods: Cinematographic recordings were performed twice in 11 healthy male participants, aged 18-25 years, without a history of spine problems, with a 2-week interval (time point T1 and T2). Image recognition software was used to identify specific patterns in the sequence of segmental contributions per individual by plotting segmental rotation of each individual segment against the cumulative rotation of segments L1 to S1. Intraindividual variability was determined by testing T1 against T2. Intraclass correlation coefficients were tested by reevaluation of 30 intervertebral sequences by a second researcher.

Results: No consistent pattern was found when studying the graphs of the cinematographic recordings during flexion. A much more consistent pattern was found during extension, especially in the last phase. It consisted of a peak in rotation in L3-L4, followed by a peak in L2-L3, and finally, in L1-L2. This pattern was present in 71% (15/21) of all recordings; 64% (7/11) of the participants had a consistent pattern at both time points. Sequence of segmental contribution was less consistent in the lumbar spine than the cervical spine, possibly caused by differences in facet orientation, intervertebral discs, overprojection of the pelvis, and muscle recruitment.

Conclusions: In 64% (7/11) of the recordings, a consistent motion pattern was found in the upper lumbar spine during the last phase of extension in asymptomatic young male participants. Physiological motion of the lumbar spine is a broad concept, influenced by multiple factors, which cannot be captured in a firm definition yet.

Trial Registration: ClinicalTrials.gov NCT03737227; https://clinicaltrials.gov/ct2/show/NCT03737227
International Registered Report Identifier (IRRID): RR2-10.2196/14741

(JMIR Biomed Eng 2023;8:e41906) doi:10.2196/41906
KEYWORDS
lumbar spine; cinematographic recordings; sequence; motion pattern; flexion; extension; rotation; physiological; musculoskeletal; motion; spine; upper lumbar; observational study; physiological motion

Introduction
Physiological motion of the lumbar spine is of interest for musculoskeletal health care professionals. Motion of the lumbar spine is dependent on multiple structures, for example facet joint orientation, spinal-pelvic relations, intervertebral disc loading, and muscle recruitment. Although the concept of physiological motion is used in many instances, a proper definition is still lacking. Over the last 90 years, several attempts to define physiological motion have been made. In 1931, Dittmar et al [1] were the first to use sagittal radiographs to analyze the normal range of flexion and extension for the lumbar spine. Subsequently, more motion research followed using other techniques including computed tomography and magnetic resonance–based 3D imaging [2-4]. Based on these data, segmental ranges of motion with a high intra, and interindividual variability were described [5,6]. For this reason, researchers started to investigate sequences, like sequence of segmental initiation of motion. Studies that report sequence of segmental initiation of motion in flexion and extension also showed variable results. The lack of consistent segmental ranges of motion or sequence hampers the definition of physiological motion of the lumbar spine [7-17].

Our research group described a consistent sequence of segmental contribution in the lower cervical spine during extension using sagittal cinematographic recordings [18]. This research was used to create a definition of physiological motion in young healthy individuals without spinal complaints. To our knowledge, similar analysis of the sequence of segmental contribution for the lumbar spine has not been carried out previously.

This study aimed to analyze the sequence of segmental contribution of L1 to S1 in sagittal rotation during flexion and extension in individual participants. A consistent pattern of segmental contribution in asymptomatic participants could be seen as a definition of psychological motion. In the future, this pattern could be used to investigate potential abnormal motion in lumbar conditions. It might be possible to better diagnose instability and the impact of it on lumbar spine motion. Furthermore, we can determine if differences in motion lead to back pain and can be resolved by physiotherapy.

Methods
Ethics Approval
The study was approved by the Medical Research Ethics Committee of Zuyderland Hospital and Zuyd University of Applied Sciences, the Netherlands (METCZ20180094).

Participant Inclusion
The study protocol was published [19]. After approval, this study included men, aged between 18 and 25 years, with a BMI <25 kg/m², with no medical history of spine problems, and able to perform maximum lumbar flexion and extension without complaints. No medical history of spine problems was defined as no visits to a doctor or physical therapist for spine complaints, no former spine surgery, total scores of Oswestry Disability Index and Visual Analogue Scale for back pain of zero, and a Kellgrens’ classification of 0-1 in levels L4-L5 and L5-S1 on cinematographic recordings evaluated by 2 spine surgeons (TB, HvS, and WvH) [20-22]. Female participants were excluded to protect their ovaries from direct radiation exposure. Potential participants were excluded if x-rays of the abdomen, pelvis, hip, lumbar, or sacral spine were taken in the previous year or in cases of active spinal infection, immature bone, lumbar tumor, previous lumbar radiotherapy, congenital lumbar spine abnormality, or planned pregnancy of the participants’ partner in the coming year. Sample size, based on previous studies, was set on 11 participants [13,14,23].

Informed consent was acquired from all participants. Radiological data were stored along with the number of participants and recordings. Handling of personal data will comply with the guidelines of the Dutch Personal Data Protection Act.

Study Procedures
Flexion and extension cinematographic recordings were acquired twice for each participant during afternoons and evenings. An interval of 2 weeks was maintained to determine reproducibility and consistency of the sequence between 2 time points (T1 and T2) [18,24]. Cinematographic recordings were made from a lateral perspective to obtain sagittal images, using the Philips Allura Xper FD20 x-ray system. The following settings were used: frames of 1024x1024 pixels, 7.5 frames per second, tube voltage of 75-90 kV, filter of 0.9 mm copper + 1 mm aluminum, and a detector distance of 48 cm. The total radiation dose for participants was categorized in category 2A, using the Neurocritical Care Society guidelines on risks of radiation dose (0.1-1.0 mSv) [25]. During cinematographic recordings, participants were seated in a customized wooden chair, designed to keep the pelvis in a fixed position (Figure 1). A 3-point fixation was located on the anterior superior iliac spine, posterior inferior iliac spine, and the upper legs, which could be adjusted to the participants’ physique. Participants were asked to remove clothes that could disturb the cinematographic recordings. From a neutral seating position with the knees in 90 degrees flexion, participants were asked to perform maximum extension, followed by maximum flexion, and then a return to maximum extension in 14 seconds, using a metronome. Maximum flexion and extension was determined as the maximum achievable position of the participant and practiced before the final cinematographic recordings. During the active motion task, arms were crossed in front of the chest (Figure 1). This duration was chosen based on the pulse frequency of the image technique (7.5 pulses per second) and the number of necessary images (104 images) for image recognition.
Figure 1. Customized wooden chair with 3-point fixation of the pelvis. The 3-point fixation is located on the anterior superior iliac spine, posterior inferior iliac spine, and the upper legs.

**Radiological Data Processing and Analysis**

For this research, we have previously developed custom software that uses image recognition algorithms to track vertebrae during flexion and extension [26]. The software follows bony structures within user-defined template areas throughout all frames, using a best-fit principle to match normalized gradient field images. To define these template areas, the user draws polygons around all vertebrae on the median frame of the recording [26]. After the software has completed tracking these structures, they can be manually evaluated. Corrections can be made if necessary. Finally, graphs are made for both flexion and extension cinematographic recordings for each individual participant to identify specific patterns in the sequence of segmental contributions. Segmental rotation of each individual segment (L1 to S1) between each pair of successive frames was plotted against the cumulative rotation in segments L1 to S1 together. A more detailed description of the image recognition software can be found in a previously published study [26]. Analyses were first performed for T1 and tested against T2. Time spent on radiological data processing and analysis was 2 to 3 days per cinematographic recording. Analyses were performed by...
researcher IC, with reevaluation of 30 intervertebral sequences by a second researcher (TB) to determine reproducibility, using a two-way mixed intraclass correlation coefficient (ICC). An ICC above 0.60 was considered adequate. A consistent motion pattern was defined as a similar pattern shown in at least 80% (8/10) of the cinematographic recordings in 2 time points. This was comparable with the results of the cervical spine [18].

Results

A total of 11 participants were recruited and included, all undergoing 2 cinematographic recordings. This resulted in a total of 22 recordings, of which 1 (P1-recording 1) was excluded from analyses, since L1 could not be followed in the field of view. No consistent pattern was found when studying the graphs of the cinematographic recordings during flexion (Multimedia Appendix 1). During extension, segments L4L5 and L5S1 showed an inconsistent pattern (Multimedia Appendix 2). Leaving L4L5 and L5S1 out of the analyses, a much more consistent pattern on the sequence of segmental contribution was found, especially in the last phase of the extension motion. It consisted of a peak in rotation in L3L4, followed by a peak in L2L3, and finally, in L1L2 (Figure 2; Multimedia Appendix 3). Only the sequence of the peaks was important, not the height of the peaks itself, since a peak represents the largest contribution of a specific segment at a specific point in the total motion despite the height. As discussed in the study of Boselie et al [18], peaks with a rotation lower than 0.3 were deemed to fall within the measurement error and were not taken into consideration. In total, 71% (15/21) of extension graphs showed the abovementioned sequence, which represents 80% (8/10) at T1 and 64% (7/11) at T2 (Multimedia Appendix 3). At both time points, P5 and P7 did not show a consistent motion sequence with different motion patterns at each time point. P9 only showed a consistent motion sequence in T1. ICC was determined for each segment in 5 cinematographic recordings (Table 1).

Figure 2. Sagittal rotation in segments in the upper lumbar spine (segments L1 to L4) during extension of the lumbar spine in healthy young male participants (P2-T1). On the y-axis, the rotation is shown in degrees between successive frames. On the x-axis, cumulative degrees of extension in block L1 to L4 are shown. Peaks of the graphs per segments (L1L2, L2L3, and L3L4) depict maximum contribution of the segment in a specific phase of the extension. At the last phase of the extension, the L3L4 peak was followed by an L2L3 peak and finally the L1L2 peak. Each series of values undergoes smoothing by means of a low-pass Gaussian digital filter.

Table 1. Intraclass correlation coefficients (ICC) per segment of 5 randomly chosen cinematographic recordings. An ICC below 0.60 is determined as inadequate and indicated in italics.

<table>
<thead>
<tr>
<th>Segments</th>
<th>Cinematographic recordings</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2-1</td>
<td>3-2</td>
</tr>
<tr>
<td>L1L2</td>
<td>0.601</td>
<td>0.858</td>
</tr>
<tr>
<td>L2L3</td>
<td>0.799</td>
<td>0.615</td>
</tr>
<tr>
<td>L3L4</td>
<td>0.577</td>
<td>0.819</td>
</tr>
<tr>
<td>L4L5</td>
<td>0.691</td>
<td>0.437</td>
</tr>
<tr>
<td>L5S1</td>
<td>0.763</td>
<td>0.258</td>
</tr>
</tbody>
</table>
Discussion

Principal Findings

The aim of this study was to ascertain the sequence of segmental contribution and to possibly understand physiological motion in sagittal rotation during maximum flexion and extension of the lumbar spine in asymptomatic male participants. Results showed a consistent pattern in 71% (15/21) of the recordings during the last phase of the extension with a peak in rotation in L3L4, followed by a peak in L2L3, and finally, in L1L2. However, this pattern was consistent in only 64% (7/11) of the recordings over the 2 time points.

Previous studies have used different imaging techniques to describe the range of motion and the sequence of initiation of motion of individual segments during flexion and extension of the lumbar spine. Dvorak et al [16], Pearcy et al [15], and Staub et al [11] described range of motion in rotation during maximal passive flexion and extension of each level. Since range of motion differed between studies and resulted in a high inter, and intraindividual variability, a more consistent method to define physiological motion was pursued. Initiation of motion was described by several previous studies. Because of limitations (eg, reporting pooled data instead of individual sequences, limited range of motion, analyses of part of the lumbar spine, and describing intervertebral rotation at specific time points or specific ranges of motion instead of between successive frames), results differ between studies with high inter and intraindividual variability [7-10,13,14,17].

In a cervical spine study [18], a more consistent sequence of segmental contribution during the end of the extension, namely in 80% of the participants in T1 and 90% in T2, was found using the same measurement method and setup as this study. The reliability, sensitivity, and specificity of this measurement method showed high scores, with a reliability, determined in Fleiss Kappa, of 0.80-0.84, average sensitivity of 90%, and average specificity of 85% [18]. We believe that these findings show that this method is accurate and reproducible to determine the sequence of segmental contribution in cervical spine. Even though the setup of this study was similar, we found less consistent motion patterns in the lumbar spine. We believe several variables between the cervical spine and lumbar spine contribute to our differences in consistency in motion patterns. These variables are as follows: facet orientation, intervertebral discs loadings, the spino-pelvic relationship, and muscle recruitment.

Cervical facet joint surfaces between C3 and T1 have a 45 degrees angle to the transverse plane [27]. In the lumbar spine, the superior articular process is medially orientated, and the inferior articular process is laterally orientated in the sagittal plane, with right angles to the transverse plane [28]. These differences in orientation result in less constrained facet joints of the lumbar spine, resulting in a greater freedom of motion, which could explain a less consistent movement compared to the cervical spine [27]. The uncinate process and uncovertebral joints, found from C3 to C7, also provide stability and mobility of the cervical spine by functioning as a guide rail during flexion and extension and limit rotation and bending, resembling a saddle joint [29]. Since these structures are not present in the lumbar spine, it could lead to less consistent motion patterns due to less constraint of the motion segments. Intervertebral discs of the cervical spine and lumbar spine are both wedged shaped with a larger anterior side of the disc compared to the posterior side [30]. In addition, both discs are elliptical shaped, with a larger cross-sectional area of lumbar intervertebral discs than cervical spine [30]. In this study, it is possible that the axial loading of the intervertebral disc is altered by fixation of the pelvis and the seating position. Nachemson et al [27] described a relative increase in intervertebral disc pressure, ascending from supine to standing to sitting position and from neutral position to flexion. Furthermore, forced anteverision or retroversion of the pelvis caused by the fixed position could influence motion patterns of the lumbar spine during flexion and extension. There is no study that compares motion of the lumbar spine in a standing versus sitting position. The pelvis and abdominal structures also led to overprojection in segment L5S1, making it challenging to trace these segments with the computer software. For this reason, ICCs were determined in this study, resulting in an average ICC of all segments mostly above 0.60, except for L5S1. Furthermore, analyses showed that the lower lumbar spine segments, L4L5 and L5S1, showed inconsistent patterns throughout all recordings. When excluding them from analyses, a more consistent pattern from L1 to L4 appeared. In addition to the difficulty due to overprojection at L5S1, the motion of L4L5 and L5S1 is influenced by more variables compared to the upper lumbar spine, leading to less consistency. The lower lumbar segments function as a kinematic transition zone from a highly mobile region (ie, upper lumbar spine) to an immobile sacroiliac region [31]. For this reason, it is also plausible that pathology mostly occurs in lower lumbar segments.

Finally, muscle recruitment differs between the cervical and lumbar spine. In the cervical spine, the range of rotations is mostly influenced by muscle recruitment, except for the end stages of the motion, which are influenced by gravity [27]. In the lumbar spine, rotation is controlled by muscle recruitments throughout the whole motion. Muscle recruitment and strength is affected by age, sex, motivation, pain, as well as muscle and joint physiology and geometry [27]. This means that mostly interindividual motion differences can be explained by differences in muscle recruitment and strength, which plays a larger part in the lumbar spine compared to the cervical spine motion. In addition, the 4 abdominal muscles (ie, rectus abdominis as well as external and internal oblique and transverse abdominal muscle) have a great influence on flexion of the lumbar spine, with an increased muscle recruitment per degree of flexion [27]. This could also be an explanation for the fact that lumbar flexion patterns are less consistent than lumbar extension patterns.

Strengths and Limitations

There are multiple strengths to this study. First, the intention of this study was to determine motion patterns of L1 to S1, instead of a selection of vertebrae by using a sufficient field of view. However, especially segment L5S1 was difficult to track due to the overprojection of the pelvis and abdominal structures. Additionally, it is possible that L5S1 also had less focus since
this segment was placed at the maximum bottom of the field of view. This resulted in a mean ICC below 0.60 for L5S1 and an inconsistent motion pattern throughout the recordings.

Second, this study described motion patterns during maximum flexion and extension of an individual instead of the usually reported fixed ranges to determine physiological motion. Maximum flexion and extension represents the lumbar motion in daily activity better, as it does not limit a person to move within a strict range. Furthermore, patients could move differently because they had to stay within a range of motion, which could influence the muscle recruitment. The downside of using maximum range motion patterns is the possibility of segments moving outside the field of view. This happened once in P1-T1 (Figure 3), resulting in the exclusion of this cinematographic recording from final analyses. However, Figure 2 shows a peak in L3L4, followed by a peak in L2L3 during the last phase of extension, comparable with the abovementioned most consistent sequence of motion.

Figure 3. Sagittal rotation in segments in the upper lumbar spine (segments L2 to L4, since L1 fell outside the field of view during extension of the lumbar spine in P1-T1). On the y-axis, the rotation is shown in degrees between successive frames. On the x-axis, cumulative degrees of extension in block L2 to L4 are shown. Peaks of the graphs per segments (L2L3 and L3L4) depict maximum contribution of the segment in a specific phase of the extension. At the end of the extension, the peak of L3L4 was followed by a peak of L2L3. Each series of values undergoes smoothing by means of a low-pass Gaussian digital filter.

Finally, since sequence of segmental contribution in the cervical spine showed consistent motion patterns in the study of Boselie et al [18], we used the same imaging technique for recordings, the same computer tracking software, and the same research team in this study [18]. Additionally, cinematographic recordings of all participants were supervised by the same team (IC and CH Christoph) and performed with the use of the same customized chair. The included participants were all male, around the same age, and with a BMI below 25 kg/m² to minimize the influence of age, sex, and body habitus on muscle recruitment and overprojection of abdominal structures. Female participants were excluded to protect their ovaries from direct radiation exposure. However, Staub et al [11], Troke et al [5], Dvorak et al [16], and Wong et al [8] showed no statistically significant difference between sexes in motion of the lumbar spine.

This study also had some limitations. First, sagittal balance parameters were not determined during this study, as femoral heads were not shown in the cinematographic recording. A fixed pelvis could influence the motion of the lumbar spine by forced anteversion or retroversion, which could have been determined using these parameters. Second, the measurement method used to develop the graphics was a time-consuming method. For this reason, the possibility of using artificial intelligence should be investigated to determine if it could lower the workload without losing reliability of the measurements. However, this would be more important for cervical spine analyses, as lumbar spine analyses using this method showed less consistent motion patterns, and therefore, it will have less clinical relevance. It could be possible that another analyzing method should be used to determine physiological motion of the lumbar spine. It has been suggested that center of rotation (COR), defined as the point around which motion segments of the lumbar spine move, could quantify the kinematic features of the lumbar spine [32]. COR of the lumbar spine was the main topic in many previous studies. However, conditions to determine COR varied between studies (eg, symptomatic and asymptomatic participants, different motion tasks, as well as before and after surgery). A current systematic review [32] is analyzing and summarizing data of these different studies to determine if COR could be used to define physiological motion of the lumbar spine. Unfortunately, results are not yet available.

Third, this study was conducted with 11 participants, resulting in 22 cinematographic recordings over 2 time points. Despite this small sample size, we believe that expansion of the study group would not have led to more conclusive results, since there were also intraindividual variabilities between the 2 time points.
besides interindividual variabilities, and previous research showed consistent results with similar group sizes.

Conclusions
This study aimed to provide physiological motion patterns of the lumbar spine based on the sequence of segmental contribution. A total of 64% (7/11) of the cinematographic recordings of asymptomatic young male participants showed a consistent pattern at both time points during the last phase of extension, with a peak in rotation in L3L4, followed by a peak in L2L3, and finally, in L1L2. Since 36% (4/11) of the cinematographic recordings did not show a consistent pattern, we believe that physiological motion of the lumbar spine is a broad concept, which cannot be stated in a firm definition using this method. Even in healthy participants, multiple factors are responsible for inconsistencies in lumbar spine motion patterns, which can be aggravated in case of lumbar pathology. For this reason and because of the time-consuming method for analysis, we believe the clinical relevance in this form will be limited, and it should not be used as a diagnostic tool to distinguish between physiological and pathological motions.

Acknowledgments
We would like to thank CH Christoph for the assistance during the cinematographic recordings. We would also like to thank Dr SN Van Laarhoven for his contribution to the customized wooden chair.

Data Availability
The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Graphs flexion of lumbar motion.
[PDF File (Adobe PDF File), 663 KB - biomedeng_v8i1e41906_app1.pdf ]

Multimedia Appendix 2
Graphs extension L1 to S1.
[PDF File (Adobe PDF File), 663 KB - biomedeng_v8i1e41906_app2.pdf ]

Multimedia Appendix 3
Graphs of extension cinematographic recordings L1 to L4; T1 and t2.
[PDF File (Adobe PDF File), 699 KB - biomedeng_v8i1e41906_app3.pdf ]

References

https://biomedeng.jmir.org/2023/1/e41906


25. NEDERLANDSE COMMISSIE VOOR STRALINGSDOSIMETRIE. Publication of the Netherlands Commission on Radiation Dosimetry 2017:1-20 [FREE Full text]


Abbreviations

COR: center of rotation
ICC: intraclass correlation coefficient
T: time point

Edited by T Leung; submitted 14.08.22; peer-reviewed by S Muraki, A Breen, Y Wu; comments to author 12.01.23; revised version received 15.01.23; accepted 31.01.23; published 20.06.23.

Please cite as:
Caelers I, Boselie T, van Hemert W, Rijkers K, De Bie R, van Santbrink H
The Variability of Lumbar Sequential Motion Patterns: Observational Study
JMIR Biomed Eng 2023;8:e41906
URL: https://biomedeng.jmir.org/2023/1/e41906
doi: 10.2196/41906
PMID:
Mixed Reality Platforms in Telehealth Delivery: Scoping Review

Hemendra Worlikar¹, BEng, MSc; Sean Coleman¹,², MB BCh BAO; Jack Kelly¹,², MB BCh BAO; Sadhbh O'Connor¹,², MB BCh BAO; Aoife Murray¹, MB BCh BAO; Terri McVeigh³, MB BCh BAO, PhD; Jennifer Doran¹, BA, LLB, MB BCh BAO; Ian McCabe¹, BSc, MSc, PhD; Derek O'Keeffe²,⁴,⁵, BEng, MB BCh BAO, MEng, MBA, PhD, MD

¹Health Innovation Via Engineering Laboratory, Cúram Science Foundation Ireland Research Centre for Medical Devices, University of Galway, Galway, Ireland
²Department of Medicine, University Hospital Galway, Galway, Ireland
³Cancer Genetics Unit, The Royal Marsden National Health Service Foundation Trust, London, United Kingdom
⁴School of Medicine, College of Medicine Nursing and Health Sciences, University of Galway, Galway, Ireland
⁵Lero, Science Foundation Ireland Centre for Software Research, University of Limerick, Limerick, Ireland

Corresponding Author:
Hemendra Worlikar, BEng, MSc
Health Innovation Via Engineering Laboratory
Cúram Science Foundation Ireland Research Centre for Medical Devices
University of Galway
Lambe Institute
Galway, H91 TK33
Ireland
Phone: 353 899413431
Email: worlikarhemendra09@gmail.com

Abstract

Background: The distinctive features of the digital reality platforms, namely augmented reality (AR), virtual reality (VR), and mixed reality (MR) have extended to medical education, training, simulation, and patient care. Furthermore, this digital reality technology seamlessly merges with information and communication technology creating an enriched telehealth ecosystem. This review provides a composite overview of the prospects of telehealth delivered using the MR platform in clinical settings.

Objective: This review identifies various clinical applications of high-fidelity digital display technology, namely AR, VR, and MR, delivered using telehealth capabilities. Next, the review focuses on the technical characteristics, hardware, and software technologies used in the composition of AR, VR, and MR in telehealth.

Methods: We conducted a scoping review using the methodological framework and reporting design using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) guidelines. Full-length articles in English were obtained from the Embase, PubMed, and Web of Science databases. The search protocol was based on the following keywords and Medical Subject Headings to obtain relevant results: “augmented reality,” “virtual reality,” “mixed-reality,” “telemedicine,” “telehealth,” and “digital health.” A predefined inclusion-exclusion criterion was developed in filtering the obtained results and the final selection of the articles, followed by data extraction and construction of the review.

Results: We identified 4407 articles, of which 320 were eligible for full-text screening. A total of 134 full-text articles were included in the review. Telerehabilitation, telementoring, teleconsultation, telemonitoring, telepsychiatry, telesurgery, and telediagnosis were the segments of the telehealth division that explored the use of AR, VR, and MR platforms. Telerehabilitation using VR was the most commonly recurring segment in the included studies. AR and MR have been mainly used for telementoring and teleconsultation. The most important technical features of digital reality technology to emerge with telehealth were virtual environment, exergaming, 3D avatars, telepresence, anchoring annotations, and first-person viewpoint. Different arrangements of technology—3D modeling and viewing tools, communication and streaming platforms, file transfer and sharing platforms, sensors, high-fidelity displays, and controllers—formed the basis of most systems.

Conclusions: This review constitutes a recent overview of the evolving digital AR and VR in various clinical applications using the telehealth setup. This combination of telehealth with AR, VR, and MR allows for remote facilitation of clinical expertise and further development of home-based treatment. This review explores the rapidly growing suite of technologies available to users within the digital health sector and examines the opportunities and challenges they present.
Introduction

Background

The term telemedicine refers to the provision of clinical health care services over a distance through information and communication technology (ICT) channels. Telemedicine overcomes geographical barriers in facilitating remote medical services. Building on this, the concept of telehealth extends to include continuing health education, research, and evaluation by medical professionals, all while promoting the health outcomes of individuals and communities [1]. Telehealth broadly encompasses the delivery of remote health-related services, including nonclinical services such as medical provider training; medical education; public health education; administrative meetings; and electronic exchange of clinical data enabling diagnosis, evaluation, consultation, treatment, and care management. The term telehealth has evolved as available technologies have improved, such that the term “digital health” is now often used as a more inclusive term reflecting the application of various different types of technologies and telecommunications systems in health care delivery. Digital health platforms can be either provider-to-provider or direct-to-consumer systems supported by the ICT infrastructure [2,3]. The telehealth sector has seen an effective increase in the past few years and has grown exponentially because of COVID-19 pandemic restrictions. According to the report published by Fortune Business Insights, the global telehealth market size was estimated at around US $144.38 billion in 2020 and is likely to reach US $636.38 billion by 2028 [4].

From the reality-virtuality continuum model, according to Milgram et al [5] (as seen in Figure 1), the real environment is that which is viewed without any overlay of the computer-generated entity, while at the opposite end of this continuum, immersive virtual reality (VR) is observed as completely enhanced computer-generated environments viewed through a head-mounted display unit. In the augmented reality (AR)–based display, digital information or entities are overlaid in the real environment, such that different aspects of reality are observed between the real and virtual environment. These augmentation-based realities can be discovered by optical see-through head-mounted displays (HMDs), mobile phones, tablets, or computer monitors [5].

Figure 1. Representation of reality-virtuality continuum by Milgram et al [5]. AR: augmented reality; AV: augmented virtuality; MR: mixed reality; VR: virtual reality.

In AR technology, the digitally created data directly coincide with the user’s real-world environment, where the user can see the computer-generated 2D or 3D entities such as holograms. The virtual entities superimposed or mapped onto the real-world space are typically rendered using optical see-through display such as HMDs or mobile-based devices, also allowing for stereoscopic visualization. The next most advanced form of reality platform, the mixed reality (MR), follows the footstep of AR and allows interaction with these virtual entities by using hand gesture inputs, gaze recognition, or controllers. The VR platform is a completely enhanced digital representation featuring a 2D or 3D virtual environment or objects that can replicate real-life surroundings. VR provides engaging sensory perceptions for both visual and acoustic stimulation. Immersive VR relies on headsets or stand-alone VR devices, whereas nonimmersive VR relies on the monitor display [6].

The introduction of VR and AR technologies in medicine has been focused on clinical-related research. The key areas incorporating this digital reality are surgery, psychology, neurological condition, rehabilitation, and medical educational [7]. The 3D picturing capabilities of the VR- and AR-based platforms have been sought for applications in the visualization of scientific experimental imaging data, tools for surgical planning and studying anatomy, and other collaborative interfaces for education and telehealth [8]. Surgical simulation has distinctively used digital reality, while VR is principally used for visual and haptic rendering, whereas AR and MR were predominantly positioned for the tracking system and graphical rendering, with the latter being used in a real surgical setting [9]. The usefulness of VR education and training using simulation methods for nursing students was comparable with the standard models of education and training on the outcomes of skills, confidence, satisfaction, and performance time [10]. The current prospects of AR software applications in medical criteria are treatment and training based [11]. Surgical development using an MR platform has been linked as a predominant utilization tool for training and simulation.
technology, advanced imaging and navigation, and broadening the extent of clinical application. Recently, MR has been adapted to neurosurgery, otolaryngology, ophthalmology, urology, and dentistry [12]. Digital reality technology has been incorporated into the preoperative surgical planning for several cranial-based applications for the neurosurgical subspecialty [13]. VR-based exposure therapy is used for various psychiatric disorders such as anxiety, trauma and stress, neurocognitive disorders, and several mental disorders. The effects of VR have been studied to have long-lasting positive outcomes for the treatment [14]. VR-based training has been effective in the improvement of executive limb function and cognitive function in patients with stroke [15,16].

Objectives

Many published studies have reviewed the use of AR and VR capabilities in medical research and practice and have not detailed its implication in telehealth, thus addressing this research gap. This systematic scoping review provides an overview of the prospects of AR and VR applications delivered using telehealth platforms in clinical settings. This review offers end users and providers an update of the current use of AR, VR, and MR effectively in telehealth delivery and highlights the prospects of such technologies in the future. This review aims to explore the following research questions:

- What clinical specialties have incorporated digital reality platforms such as AR, VR, or MR exclusively with telehealth?
- What are the different hardware and software technology formats used in AR, VR, or MR within telehealth?
- Which important technical features of AR and VR have been used in telehealth?

Methods

Overview

This scoping review used the framework of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [17]. Included studies from the database were solely concerned with the application of high-fidelity simulation technology such as AR, VR, or MR exclusively delivered via the telehealth platform. The study has no written or published protocol.

Database and Search Strategy

Articles from Embase, PubMed, and Web of Science were explored to obtain relative pieces of evidence. An exploded search strategy string was developed with the support of a university librarian. The search string included appropriate keywords and Medical Subject Headings terms—“augmented reality,” “virtual reality,” “mixed reality,” “extended reality,” “telemedicine,” “telehealth,” “m-Health,” “e-Health,” and “digital health.” The search strategy was initially developed on the Embase database and replicated across the other databases using predefined filtering techniques. The entire search strategy can be seen in Multimedia Appendix 1.

Eligibility Criteria

The studies included must satisfy the active component use of AR, VR, or MR delivered via telehealth approaches and should have been published between the years 2016 and 2021, since such devices with this technology format became commercially available, marked in reference to the release date for the first-generation Microsoft HoloLens [18]. The collaboration aspect of AR and MR technology into social or digital communication avenues could be observed during the same period [19]. Telemedicine or telehealth includes a broad spectrum of health care delivery, including education prospects; however, this review will focus on clinical aspects, including simulation. Only full-length text articles available on the web in the English language were included. Full-length text from peer-reviewed articles such as randomized controlled trials, feasibility studies, exploratory studies, narrative reviews, systematic reviews, case and cohort studies, book sections, and technical reports was considered eligible for inclusion. Any studies highlighting the mentioned technology for gaming, entertainment, or medical education were excluded. Correspondence papers, letters, conference abstracts (no full texts), editorial, commentary, poster presentations, and gray literature were also excluded from this review.

Study Selection and Data Extraction

The papers obtained from the applied search strategy from the information databases were imported to the reference manager EndNote 20 library, and duplicates were discarded [20]. Three researchers (HW, SC, and JK) performed initial screenings based on titles, abstracts, and keyword searches. Author HW conducted eligibility criteria and full-text screening. The selected studies were then reviewed based on the article type, study design, clinical condition addressed in the study, mode of telehealth communication, acceptance criteria, and the hardware and software used in the studies for the guidance for data synthesis. Finally, the relevant information from the studies was tabulated into an Excel (Microsoft Corp) spreadsheet, and a descriptive synthesis of the data was generated. In our review, we summarized and grouped the various telehealth branches using digital reality platforms for the various clinical condition based on descriptive statistical findings for the included studies. Different facets of the digital reality technology were detailed for its application in clinical research.

Results

Overview

Of the 4407 abstracts identified from the search protocol, 134 full-text articles fulfilled the inclusion criteria. A total of 1079 duplicate records were removed, 2598 records were discarded after title, abstract, and keyword search, and 410 records were deemed not fit after the initial screening as these articles were not about topic of interest having objectives that did not align with the outcomes of this review and did not satisfy the inclusion criteria. Of the 320 articles that were subjected to full-text review, 177 articles were deemed not relevant because they either included the digital reality technology or telehealth strategies but not delivered jointly, and 9 were excluded after
recognizing multiple papers published on the same topic by the HW, SC, and JK (Figure 2).

**Figure 2.** Flowchart for the structured literature search and selection.

**Digital Reality Platform via Telehealth**

As demonstrated in Figure 3, VR and AR cover most of the listed telehealth domains for the eligible studies. The most studied and researched area is telerehabilitation accomplished using VR. The other subareas involving VR use include telepsychiatry for evaluation and treatment, telediagnosis, and teleconsultation. In addition, AR and MR are prevalent modes of the reality technology platform for telementoring and teleconsultation. Finally, telesurgery and telemonitoring are the 2 subfields of telehealth where AR technology have seen an upward trend.

Clinically based digital health applications were considered for the review as various specific branches of the telehealth spectrum (Figure 4). Telerehabilitation is a postclinical care service delivered at home or remotely for recovery purposes and constitutes most of the included telehealth group from the included studies [21,22]. Evident from the included studies, stroke rehabilitation emerges as the leading medical condition.
that has seen an uptake of these services. Different aspects of rehabilitation, such as functional motor training, including upper-extremity training and fine motor skills, cognitive functional training, visuomotor tracking training, and balance and gait training are primarily used for treating poststroke survivors [23-41]. In turn, the patient groups who have used telehealth for the purposes of rehabilitation have reported improvements in their quality of life, increased daily activities, and improved levels of motivation [42]. From the multiple studies included, telerehabilitation has been experimented as a home-based treatment for various neurological and cognitive disorders or diseases such as Parkinson disease, acquired brain injuries, multiple sclerosis, cerebral palsy, mild cognitive impairment, Alzheimer disease, and dementia [43-58].

Conventional therapy programs in the form of physical therapy and behavioral therapy are the nonpharmacological treatments that have used this remote delivery platform. In a small number of studies, the home-based rehabilitation in the form of novel telerehabilitation have been used for patients undergoing surgical procedures, such as total hip replacement, total knee arthroplasty, and total knee replacement, as a postrecovery treatment measure [59-62]. Mirror therapy for patients with phantom limb pain and physiotherapy treatment for patients with chronic body pain have incorporated this model of remote teletherapy [63-68]. This field has also been applied in physical rehabilitation for musculoskeletal disorders, provision of vestibular rehabilitation therapy in patients with a balance disorder, and kinesiotherapy for older adults at risk of falls [2,69,70]. Physical therapy in the pediatric group and musical therapy in patients with spinal cord injury have explored this stream of technology [71,72]. Pulmonary rehabilitation therapy for respiratory disorders such as chronic pulmonary respiratory disorder, pulmonary fibrosis, and myocardial infarction; low vision rehabilitation in providing functional visual assistance; and the COVID-19 pandemic have been an influential factor in accelerating remote rehabilitation therapy [22,73-77].

Telementoring is a subset of telemedicine that reflects remote expert guidance such as training or telenavigation to medical and nonmedical personnel in performance of life-sustaining procedures [78]. The impact and usability of the telementoring technique in provision of cardiopulmonary resuscitation in treating cardiac arrest has been demonstrated by different authors in simulated environments, with the assistance of a remote mentor using an HMD or Google Glass [79,80]. Other authors have explored the use of telementoring guidance, in intraoperative telenavigation, and preoperative planning in simulated battlefield and emergency trauma. The telementoring approach for preoperative planning and telenavigation during the intraoperative process has been demonstrated in complex emergency hand reconstruction surgery [81,82]. Forward damage control procedure performed on a patient-simulator model depicting a right-sided femoral gunshot wound and simulated trauma injuries such as airway obstruction by conducting cricothyroidotomy have been carried out using remote instruction—as have, lung decompression, tracheostomy, or REBOA (resuscitative endovascular balloon occlusion of the aorta) catheter deployment to deal with specific trauma injuries with the aid of a remote medical expert [82-86]. The feasibility of telementoring applicability in the performance of chest thoracotomy, skin grafting, and fasciotomy has been evaluated using ex vivo animal models [78,87,88]. Telementoring has been used to great effect in different stages of surgical planning in various orthopedic, craniofacial, spinal cord, vascular, and cardiothoracic surgeries [6,89-100].

**Figure 3.** Collaboration between digital reality technology and telehealth for the included studies.
Teleconsultation is a primary segment of telehealth services, broadly consisting of remote consultation services using ICT. This remote consultation can be synchronous or asynchronous and between clinicians (provider-to-provider) for shared decision-making or between clinicians and their patients (provider-patient) [101]. This approach has been applied in patient assessment using the National Institute of Health Stroke Scale for patients with acute stroke and during remote clinical rounds in isolation wards for patients with COVID-19, thereby reducing direct exposure of the staff [102,103]. This technique has also been evaluated in trauma and emergency-related scenarios, such as remote consultation in reading and interpreting electrocardiogram reports related to drug intoxication or poisoning [104,105]. The effectiveness of provider-to-provider teleconsultation has been demonstrated in provision of support for ambulatory staff and first-responders in triage during simulation of major trauma [106]. The applications of teleconsultation in provision of surgical care are broad, allowing collaborative, contextual, and presurgical planning and visualization and intraoperative surgical navigation through high-fidelity immersive reality platforms and devices, as well as facilitating remote delivery of complex information to patients [107-115]. Teleconsultation via the reality platforms has been used to explore the feasibility of telepathology in carrying out an autopsy, image scanning, and transfer of serially sectioned cancer tissue from a mouse [116].

Telemonitoring is an advanced form of clinical care service that provides patient-centered care. This method allows health care providers to collect and track patient information and deliver remote care assistance [117]. This branch of telehealth has been evaluated in pediatric cohorts dealing with hospital-induced stress as a shared experience on a mobile-based AR game for
play therapy. This aspect allows managing pediatric patient profiles, data collection, and further analysis for effective treatment [118]. Telemonitoring via haptic conversational agents; that is, a computer-generated character to deliver physiotherapy home exercises to patients with musculoskeletal disorders and chronic pain has been demonstrated to increase their treatment adherence [119]. Supervised AR-based home training has been used for patients with phantom limb pain by providing mirror therapy, thereby promoting visuomotor integration by reengaging the neural circuits related to lost limbs [120]. Telemonitoring has been used in postoperative care and wound assessment in orthopedic and neurological cases and has also been applied for teleproctoring or remote monitoring in pilot simulation as training for fundamentals of laparoscopic surgery examination [121-123].

Telepsychiatry uses ICT to offer a range of clinical and nonclinical services such as psychiatric evaluation, therapy (individual or group-based), patient education, and management remotely [124]. Studies using this element of telehealth and computer-generated virtual environments have evaluated the feasibility of remote therapy such as Virtual Reality Exposure Therapy for patients with acrophobia and evaluation of the technical system in delivering specific phobia treatment for arachnophobia [125,126]. Remotely delivered psychological treatment by the mental health professional include behavioral intervention therapy, cognitive behavioral therapy, mindfulness therapy, and acceptance and commitment therapy for patients facing stress, anxiety, public speaking anxiety, and social anxiety disorder, among others [127-130]. In a simulation study, cognitive and affective assessment of astronauts has been carried out to characterize social isolation from space [131]. Evaluation of telepsychiatry using the reality platform such as VR versus the traditional videoconferencing platform, and the development of newer platforms such as social VR for older adults in urban areas has demonstrated such techniques could lead to improved quality of life by reducing social isolation [132]. Telepsychiatry assessment via VR as a home-based treatment delivered by mental health professionals, such as a psychiatrist, psychologist, licensed social worker, or a mental health counselor, has been demonstrated to mitigate clinician burnout [133,134].

Another exciting subsection of the telehealth sphere, telesurgery, enables teleoperation in an operating field executed over a distance. Telesurgery involves using various disciplines such as communication technology, imaging techniques, motor control systems, robotics, reality platforms, and digital signal processing [135,136]. For example, in an experimental setup, a VR-based teleoperative system consisting of a robotic catheter operating system can be used to imitate vascular interventional surgery for arterial aneurysms or other vascular diseases. This method allowed unskilled surgeons to train in essential catheter guidance skills and enabled experienced physicians to conduct surgeries cooperatively [137]. In addition, a telesurgical experiment was conducted with a tendon-driven continuum robot via telenavigation for endoscopic and minimally invasive surgical procedures by tracking coordinate trajectory registration [138]. Finally, in another simulation case, a magnetically driven endoscopic capsule enabled the teleoperator or user to receive visual feedback in VR to conduct capsule endoscopy for colorectal cancer [139].

Moreover, the reality platform is streamed as a functional stereoscopic display and navigates space during telesurgery. This aspect of telesurgery has been experimented with as a visualization opportunity using smartphone-delivered vision and VR headsets to perform microsurgery for cataract and phacoemulsification [140]. In addition, Stereoscopic AR Predictive Display using the da Vinci R Surgical System to perform laparoscopic surgery and AR-assisted robotic surgery for kidney transplant procedures are some of the current practical applications of telesurgery [141-143]. Telediagnosis refers to the detection or evaluation of a disease or condition using telematics technology. It is achieved remotely while the patient is at a local site with remote diagnostic tools and devices [144,145]. For instance, in experimental analysis, locating and evaluating tumor-bearing hysterectomy coordinates during a 3D navigated gynecological operation facilitates telediagnosis when visualized on a 3D user interface of the medical record [146]. Another study proposes a framework based on bidirectional haptic feedback and tele immersion in the evaluation of range of motion and maximum isometric strength using the 10 arm movements method in the diagnosis of musculoskeletal disorders, poststroke rehabilitation, or postshoulder surgery [147]. Ultrasoundography (USG) is a field in which telediagnosis using the high-fidelity visualization system has been used to great effect. Evaluation of 3D VR telenavigation in cardiac USG has been undertaken in simulated settings. The added benefits of AR enable real-time telediagnosis on procedural performance and image registration for point-of-focus ultrasoundography (POCUS) and foveated imaging pipeline in extending VR-based telediagnosis [148-150]. Another study mentions AR video communication projected by mobile-based AR guidance to conduct POCUS on popliteal nerve block and a subsequent diagnosis based on the availed health information [151].

Overview of the Hardware and Software Units for the Included Studies

To experience MR, high-simulation visualization hardware devices and some of the commercial ones included in the selected studies are listed in Multimedia Appendix 2. These include high-end AR and VR devices, smart glasses, mobile devices, standard LED (light emitting diode) and LCD (liquid crystal display) television or display screen, 3D television, and 3D projectors. The commonly included immersive reality-capable devices are mostly wearable technology such as smart glasses, VR or AR HMDs, and nonimmersive standard display units. However, these high-fidelity simulation display technologies form the final part of any system and are primarily used in combination with optical capturing and tracking devices and input devices. The optical capturing and tracking systems or devices incorporate 3D depth and color-sensing camera sensors. The input devices such as controllers, trackers, or customized input modules help navigate the immediate VR or any MR environment. Various studies have included the VR gaming element in their rehabilitation programs, with some having their own developed VR rehabilitation system. Most included studies have used biometric devices for specific

https://biomedeng.jmir.org/2023/1/e42709

JMIC Biomed Eng 2023 | vol. 8 | e42709 | p.31

(page number not for citation purposes)
medical parameter evaluation to draw analysis and simulation models to conduct various training. Other relative hardware devices and systems that have been used are listed in the Multimedia Appendix 3.

The graphical representation for any software harbors a visualization platform, and more specifically, the MR system incorporates contextual 3D figures and scenes. The various software applications and source platforms that were featured and used in the included studies are listed as a table in the Multimedia Appendix 4. These applications are grouped as 3D modeling and visualization software, communication and streaming software, file-sharing and transfer applications, and other specific and personalized software applications. The 3D composite images and environment for the MR technology are created using the computer graphic designing software and gaming engine platforms. Processing and accessing the 3D computer-generated environment or images needs specific and compatible visualization file applications supported by the device. The featured 3D modeling and viewing application allows for creating and editing static and interactive multidimensional models and VR scenes, animations, and games, conversion of the produced scanned images to computer-aided design models, and stereoscopic 3D display content. The telehealth domain explores the ICT for effective remote clinical services while using the streaming facilities offered by various low-bandwidth platforms. This domain allows offline or real-time interactive communication and collaboration for any dedicated clinical services. Many communication and streaming applications allow for remote one-to-one or group video calls and messaging, screen sharing, file sharing, hosting channels, and video broadcasting. Some of these platforms allow for direct AR and VR integration and acceptance. The file-sharing and other specific applications are synchronously and explicitly used as a sequential fragment of the entire system. The developed software from the studies mentioned in the table encapsulates the combination of AR and VR cooperatively with the remote telehealth applications.

**Virtual Environment**

A virtual environment (VE) recreates a coordinated appearance of sensory information representing that of a physical environment that can be unreal, interactive, or wholly imagined environment perceived when the user wears an appropriate gadget [152]. In addition, the term virtualworlds has been interchangeably used with a VE. Developing this state-of-the-art perceived environment is created using a subset of tools arising from computer game technology, specifically through commercial game engines. The scene can be a 2D or 3D illustration, which is a complex and time-consuming process for its creation [153]. This element of VE has been recreated in almost every aspect using the VR platform. For example, the study by Levy et al [125] demonstrated the use of virtual worlds such as a subway station and a 24-story high-rise building as background scenes to overcome acrophobia as a VR exposure therapy. Similarly, Cikajlo et al [127] developed a program called ReCoVR (Realizing Collaborative Virtual Reality for Well-being and Self-Healing). The participant attends a remote guided mindfulness program as part of a group. This mindfulness program was organized as 360° video scenes where they carried out different tasks and exercises. Initially, all the participants that joined were seated in the virtual fireplace room; upon the program’s progression, they were switched to other 3D VE s, such as the Dooney Rock, River Bonnet, or the mountain-view. Shao and Lee [132] have addressed a social VR platform that uses the 3D scenes in the VE for real-time face-to-face communication in different distant locations to learn about its value and urgency in the urban older adult population. Tamplin et al [72] developed a web-based music therapy telehealth platform using social VR, vTIME (vTime Limited), allowing group music therapy sessions in VE, such as singing around a campfire in a forest.

**Gaming-Based VE**

Moreover, many studies used VE in interactive game–based settings for rehabilitative exercise programs. In a program described by Meca-Lallana et al [53], patients were required to carry out specific tasks to accomplish a mission in 2 different scenes: a medieval fantasy world and a deserted island. Yet again, in another exercise setup, VR exercises depict a wooden church in Hrabova Roztoka. The patient explores this particular place using a VR headset, thereby facilitating lower-limb rehabilitation [36]. Telehab VR, a custom-built application program that runs on either a mobile-based tablet or PC, was developed using the game engine Unity (Unity Technologies Inc). This system provides upper-limb rehabilitation for patients with multiple sclerosis. They perform various activities of daily living tasks happening in the VE in a realistic home setting. A leap motion controller (Ultraleap) was used to track and control the hand motion executed while performing the gaming tasks [48].

**Telepresence**

Telepresence describes the characteristic of directly interacting with the actual physical state, experienced from the first-person viewpoint of the user located remotely [154]. Tian et al [147] used the H-TIME (Haptic Enable Tele-Immersion Musculoskeletal Examination) set up at both the patient and doctor ends to conduct a remote diagnosis of musculoskeletal examination. At both sites, the doctor and patient could feel each other’s movements because of the bidirectional force feedback mechanism. They could view and communicate with each other in the VE, bringing them to the same examination room virtually. In another instance, in treating phobia, in particular, fear of spiders, the patients were allowed to interact in the VE, where the therapist gradually added the feared creature to the scene. This treatment is performed remotely via the tactile internet with VR headsets or standard computer screens using a hand-tracking and haptic device such as a glove [126].

**Teleoperation** refers to performing designated highly skilled manual tasks remotely, similar to a telerobotic medical system in minimally invasive surgery [155]. In a simulated study, an endoscopist performs a teleoperation process using a haptic device that controls the position of an external permanent magnet positioned at the end of a robotic arm. The user is wearing a VR headset and receives the corresponding visual information from the camera of the endoscopic capsule and then proceeds with the navigation process inside the colon [139].
Another simulated study used teleoperated ultrasonography that builds on the VE developed as a 3D representation of a real USG probe and a mannequin imitating a patient’s body highlighted with a geometric mesh for the purpose of following the examination. The user wears an Oculus Rift DK2 HMD (Oculus) to perform this simulation of tele-USG [148]. Syawaludin et al [150] introduced the use of 360° foveated pipeline imaging viewed via an HMD. The image or video capture is facilitated by the use of an omnidirectional pantilt-zoom camera module, and the remote physician can remotely diagnose the wound by zooming in and inspecting it in a 360° view over the HMD.

**Exergaming and Serious Gaming via VR**

In the context of virtual telerehabilitation, exergaming and serious games are the 2 most popular applications that emerge. Simply put, exergaming is an activity connected with playing video games that involve physical exercise [70]. In contrast, serious games follow the objective of games, implicitly focusing on increasing skills and abilities and gaining experience and knowledge [156]. The TELEKIN system, a beta edition, uses the interface of the serious game to rehabilitate cognitive and musculoskeletal disorders using a web-based framework [43]. The training sessions are conducted and played in a 3D VE that includes a number of games. Two of them involve physical actions—flexion and extension of wrist, hip, or shoulder as the doctor chooses to control the ball and platform in executing the game. Another game focuses on cognitive rehabilitation by featuring random words that the user must arrange them to form a sentence, which can be achieved using gesture-based controls [43]. Gandolfi et al [44] used the Tele Wii Lab platform as a home-based balance training, and Sheehy et al [28] and Allegue et al [30] used the Jintronix system (Jintronix, Inc) in upper-extremity rehabilitation of chronic poststroke patients, which used the exergaming platform. The study by Triandafilou et al [24] that developed a networked multiplayer gaming format, Virtual Environment for Rehabilitative Gaming Exercise (VERGE), conducted a feasibility trial to determine the effectiveness of this developed system with other potential home treatments. The VERGE system features a set of 3 exercises, namely Ball Bump, where the users pass the ball back and forth across the table; Food Fight, where the users in multiplayer can pick up the food on the table and throw it at each other; and the Trajectory Trace game, where 1 player draws a trajectory path in the space while another player retraces the trajectory to erase it. Burdea et al [32] included a commercial rehabilitation system with a novel therapeutic game controller, BrightBrainer (Bright Cloud International Corp). This system offers a multitude of interactive games (Breakout 3D and Card Island Towers of Hanoi, among others), training motor, cognitive, and executive functions for chronic poststroke patients. Qiu et al [35] demonstrated the feasibility of a home-based VR system that features 12 developed games focusing on the elbow-shoulder, hand, wrist, and entire arm for upper-extremity rehabilitation in poststroke patients (finger games: car, bowling, and piano; hand games: piano and fruit picking; wrist games: Wakamole and wrist flying; and finally, the shoulder-elbow games: the Maze, Arm Flying, Brick Break, and soccer goalie).

**Avatar Representations or Virtual Agents**

The term *avatar* is a distinguishable digital characterization of a human form (either specific or random) [128]. Moreover, these avatars can be either in 2D or 3D illustration, representing a specific part of the body, usually arms or an entire body structure with particular facial expressions. 3D avatars have been a central representation in the scope of VR and AR. The study by Anton et al [59] implements the Kinect-based Telerhabilitation (ie, KiRES) interface, providing two 3D avatars to guide the patient during their physical therapy session. One of the avatars represents the remote therapist and represents the local user or the patient, colored red and blue, respectively, so the patient can follow and perform the exercises executed by the 3D remote therapist avatar (in red). The patient can see their movements reflected by their blue avatar changing their positions as per the scenes from the therapy. In the study trial conducted by Jung et al [74], in a telerehabilitation program—Pulmonary Rehabilitation in Virtual Reality (PR in VR) program—each patient was provided with a VR headset, pico G2 4k (Pico Immersive Pte Ltd), preloaded with the PR in VR application. This application contains education and rehabilitation modules, and the chronic obstructive pulmonary disorder rehabilitation module comprises several physical exercises directed by a virtual instructor in 3D avatar embodiment.

The REWIRE autonomous telerehabilitation platform offers home-based intensive rehabilitation as offline remote monitoring by hospital clinicians. This system features a virtual therapist with artificial intelligence implanted and provides real-time feedback to maintain correct posture. In addition, the exercises performed by the patients are showcased as a 3D avatar on the screen. This intelligent system highlights each body segment of the exercise in a different color, intense green for the proper posture and red color for the incorrect posture [23]. The VERGE system enables the use of avatars to control and manipulate objects in the virtual gaming environment, allowing the capability to include multiple avatars and different users to manipulate the same object [24]. In the social VR app, vTIME (vTime Limited), an avatar persona is used for self-characterization to communicate in an immersive VE [128]. Afyouni et al [65] describe the use of RehaBot, a virtual assistant that illustrates to the user how to perform the exercises correctly (both the therapist and patient can replay the session in a 3D avatar). The RehaBot embeds real-time pattern and gesture recognition together with a dynamic correction module that considers the game difficulty level and reading from the virtual assistant to produce a tailored set of exercises that are rather fitting to the patient’s native abilities.

**Telestration and Annotations via AR and MR**

*Telestration* enables the drawing of freehand representations, also known as annotations (such as lines, circles, or any other symbols or sketches) over any image or video feed. With the latest AR and MR technology, this telestration can be achieved in 2D and 3D and superimpose this annotation in the live video streaming during the video call [157]. The Virtual Interactive Presence and AR tool is a mobile or tablet-based augmented reality platform running on an iPad device (Apple). It
incorporates the telesstration feature, allowing the remote expert surgeon to freeze the screen and then draw an image using a 2D pen tool. This composite video feed, viewed on both the local and remote stations, enables intraoperative telecollaboration in real-time [90]. In the feasibility study by Wang et al [100], POCUS using the HoloLens was conducted by the trainee in a simulated teleconference session. The MR capture video from the trainee was broadcasted. Live guidance provided by the expert mentor facilitates the trainee to complete a right quadrant Focused Assessment using Sonography in Trauma examination. The broadcasting was achieved using VSee, a proprietary low-bandwidth, group videoconferencing and screen-sharing application. To perform complex hand reconstruction of a patient after a bomb-blast injury, a telementoring network was established between an expert surgeon in Lebanon and a local surgeon in Gaza. This session was hosted using a cloud-based AR platform Proximie (Proximie Limited), allowing the remote surgeon to superimpose their own hands or range of annotations and drawing tools into the virtual surgical field [82].

Mitsuno et al [93] demonstrated telementoring in a simulation study to perform craniofacial surgery by using a teleconferencing setup, Skype (Microsoft Corp) for HoloLens, enabling the teleseatured features and images overlaid on the receivers’ visual field. A POCUS examination was performed using a novel smartphone app Vuforia Chalk (PTC Inc), an AR video platform for remote AR assistance, anchoring the AR annotations in each other’s supposed visual environment [151]. In the study, Ritcher et al [141] proposed the first predictive display with AR registration and rendering using stereoscopic displays designed for teleoperated surgical robots known as Stereoscopic AR Predictive Display. The simulation study measured the effectiveness of Stereoscopic AR Predictive Display conducted on the da Vinci R Surgical System (Intuitive Surgical) to complete the peg transfer task. The System for Telementoring with AR (STAR) platform now combines optical see-through display, HoloLens AR HMD. Similarly, this system allows for telementoring guidance by overlaying 3D graphical annotations onto the mentee’s view of the surgical field, which remains anchored in the same place even after the mentee moves their head position [95].

The experiment by Zhang et al [110] aimed to enhance teleconsultation by using the AR technology ARKit (Apple) to create an immersive replica of the consultant. Using a Kinect sensor (Microsoft Corp) to capture the skeletal feature points of the consultant, the patient views a 3D dynamic virtual avatar doctor appearing in the patient’s telepresence environment on their iPad device. A qualitative study was conducted to gain the experience and perception of AR Glasses in patients with pulmonary disorders for home-based televelitation. The web-based televation system Optimov (Optimov) enabled via an AR Glasses device Laster WAV headset (Laster Technologies) provides exercise coaching using a 3D virtual agent [73]. A holographic virtual therapist was deployed in the HOLOBALANCE, a novel health care platform for providing vestibular rehabilitation therapy for patients with balance disorders [158]. In the design and evaluation user study by Kowatsch et al [119], a hybrid ubiquitous coaching model relying on mobile and holographic conversation agents was introduced. The 3D virtual conversation agent demonstrated the squat exercise, engaging in real-time audio feedback for counting the repetition or providing automatic error detection for incorrectly or incompletely following the exercise. An innovative 3D point tracking module and unique AR system integrated with the HoloLens was used for surgical applications using telementoring. This module allowed for real-time 3D position tracking of the virtual scalpel handled by an experienced surgeon remotely. The inexperienced trainee wearing the HoloLens can see the surgical annotation superimposed with the actual surgical scene; the virtual path coregistered on the phantom arm model [78]. Next-generation mobile-based AR games for pediatric health care allow shared experiences with multiple other AR-supported devices to detect and interact using the same local area network. Several games were developed using the Unity game engine and ARCore Unity, a software development kit for Android operating software. Jungle Adventure, Map explorer, and Wakamole implemented AR interaction, whereas Map explorer and Wakamole particularly enabled the inclusion of a 2-player for a shared collaborative experience [118].

First-Person View for AR Capture Video Feed

Noorian et al [102] demonstrated smart reality glasses to conduct remote consultation using the National Institutes of Health Stroke Scale scores for stroke assessment. The onsite doctor wears the reality glass, Google glass. This Google glass is embedded with the Xpert Eye platform (AMA XpertEye), capable of assisted reality, allowing the person wearing this device to share their field of view in a 2-way real-time videoconferencing. Similarly, Nikouline et al [123] presented a feasibility study using the Google glass live video stream coming from the onsite proctor and the participant tasks related to fundamentals of Laparoscopy for scoring and evaluation done by the remote proctor. In their experiment, Lin et al [84] implement projective video texture-mapping that supplements a robust high-level stabilization video feed obtained from the mentee’s first-person view. This effective format provides the remote expert with an effective workspace visualization, allowing seamless integration of annotations in an effective AR surgical telementoring. The prospective observational study by Martin et al [103] uses HoloLens 2 MR device to conduct remote clinical consultation in a COVID-19 ward. A senior staff member would enter the COVID-19 ward to undertake clinical rounds, and the other staff members of the staff team would join virtually, thereby minimizing exposure and infection transmission. Dynamic 365 Remote Assist (Microsoft Corp) software allowed for bidirectional audio and video functionality through which the remote staff team could see the first-person view from the HoloLens 2 device. In addition, this platform allowed to place relevant imaging and electronic health record data in the user field of view, improving situational awareness and better clinical decision-making. Finally, it significantly reduced the risk of direct viral transmission.

Web- and Cloud-Based Telehealth Delivery Modes

As digital communication network and services evolve, these are rapidly being adopted in health care delivery. The web- and
cloud-based applications have become prevalent in telehealth. Telehealth relies on the backbone of internet infrastructure supported using various broadband connections such as digital subscriber lines, fiber broadband, and wireless connection, including fixed wireless broadband, cellular network or mobile broadband, and satellite communication. Thus, ICT has become central to offering an array of digital health solutions such as real-time audio and videoconferencing, remote patient monitoring, store and forward technologies, and mobile health, among others [159,160]. A web-based application principally operates on the webserver. It is accessed through a web browser over an internet connection, whereas cloud-based applications operate similarly to web applications, operating on either or both the client and server sides [161]. The custom-developed systems KiRES [59] and STAR [85] rely on the WebRTC framework, an open-source application programming interface allowing for real-time audio-video and multimedia connection. In addition, the study by da Silva et al [55], included a web-based gaming application MoveHero, to evaluate the feasibility of home-based nonimmersive serious games in patients with cerebral palsy.

Interestingly, any virtual web-based application feeds on the information; thus, data storage and hosting become integral to all online services. The study by Kato et al [42], adopted the cloud-based storage and file hosting service Dropbox (Dropbox Inc) for collecting the spatial coordinate data for each joint using the 3D optical camera during the VR telerehabilitation. The proof-of-concept study by Sirilk et al [107], implemented an e-consultation system based on the AR and MR systems using the HoloLens device for remote consultancy services in the intensive care environment. This e-consultation platform depended on a cloud-based data center that performed as an information exchange and provided services for the end devices. It also consisted of body area network technology to integrate the vital physiological information from different client devices to the data center. Prvu Bettger et al [62], used a virtual telehealth system—virtual exercise rehabilitation assistant or virtual exercise rehabilitation assistant (Reflexion Health, Inc)—for posthospital care for total knee arthroplasty. Tsouiris et al [2] included a custom-developed platform HOLOBALANCE system in managing balance disorders, both using the technology-forward cloud-based platform.

**Discussion**

**Principal Findings**

This scoping review explores state-of-the-art extended reality platforms and telehealth solutions used in the clinical context. This review highlights the reported evidence-based practical and probable applications of the extended and MR platform with telehealth used in different clinical specialties. This review also addresses the technical characteristics of the AR and VR features used in telehealth services, including various hardware and software arrangements.

Stroke is the leading clinical condition incorporating telerehabilitation, a segment of the telehealth service and digital VR [23-41]. Approximately half of the included studies from the search strategy feature the use of telerehabilitation (Figure 4). Other clinical conditions such as neurological or cognitive disorders, musculoskeletal conditions, and postsurgical recovery have also adopted telerehabilitation facilities in the home or remote settings to continue treatment. Telerehabilitation used technical attributes of exergaming and serious gaming in improving the motor and cognitive functional skills [43-58].

Other divisions of telehealth, namely telementoring, teleconsultation, and telemonitoring, have been more frequently used for surgical-related procedures, emergencies and trauma, and in several disaster simulation for disaster response and preparedness [162]. The AR and MR technologies are more prevalent with telementoring, teleconsultation, and telesurgery (Figure 3). Exposure therapy under telepsychiatry has used VR to give the patient a photorealistic experience of overcoming their pathological response to their fear [126].

Telestrated AR features through anchoring of annotations in real-time and space, performed remotely via various communication channels: a useful aid in telesurgery [81]. The technical features from the digital reality technology of VE, digital avatars, telestration or the 3D rendering of annotations, and first-person viewpoint have demonstrated teledemographic capabilities. The web- and cloud-based applications have various potential uses across the web-based clinical sphere [110]. Most of the included studies relied on existing commercial high-fidelity simulation technological hardware devices such as head-mounted AR and VR displays. The study and software designs for most of the included studies were codeveloped by the respective research teams by using multiple supportive platforms as a direct requirement for the project objectives.

These novel reality technologies of AR, VR, and MR enable 3D visualization, thereby creating a visual sense and experience of high ecological validity [163]. This technology has been extended as a remote, home-based solution for patients, thereby enabling patient empowerment [164]. This technology is highly engaging and motivating from the patient responses to telerehabilitation, consequently necessitating initial patient training needs may become an arduous task to the facilitator [42,165]. Network connectivity, internet and server security concerns, and technological constraints are some of the most common pitfalls across several studies included in this review [99,166]. The lack of interoperability between the hardware and software platforms poses a significant challenge in realizing the potential of this technology [2]. The need for improved network infrastructure and scalability poses a challenge and target for telehealth services; however, there is a risk of complete network failure, which can affect the use of such systems in critical care applications [83]. Patient confidentiality is integral at any stage during electronic exchange of health-related data; thus, network and data security protection are crucial factors for accessing telehealth services and should be robustly adhered to the governing regulations [66,149,167].

The exploded search strategy captures a broad array of important clinical applications of this high-fidelity reality technology and telehealth facilities. This review presents a current road map and the prospects of digital reality technology and telehealth in the clinical space. The determining factors presented allow the readers and researchers to evaluate the relevance of this
technology and its subsequent uptake in the clinical health sector. The study protocol was not registered, the included studies were not classified for risk of bias assessment, and the general characterization for the included studies were not presented. In addition, the review only included studies available in the English language and no relevant additional pieces of information was considered from the gray literature. From the broad array of literature-based evidence, most of the included experimental studies were pilot feasibility studies with small sample sizes, leading to reporting bias.

Conclusions
This review uniquely details the current and potential applications of digital reality technologies such as VR and AR and telehealth solutions. The feasible and practical application of AR and VR in the digital clinical space has been explored, as well as the challenges this multiparty technology endures in effective implementation and adoption. This suite of technologies offers a collaborative experience among health care professionals and their patient community. The telehealth component with the high-fidelity digital reality allows for an immersive and integrative means for teleconsultation, telesurgical procedures, and telementoring among the medical peer-to-peer group allowing for effective decision-making and treatment approaches. The uptake of VR and exergaming in various telerehabilitation programs has opened new avenues to posttreatment measures. This essential application of telehealth enhances the traditional health care delivery approach by enabling remotely delivered clinical care and services and developing home-based treatment programs. Further validated studies are needed to evaluate the overall assessment of this trending technology, thereby leading to commercial pathways. A robust and secure communication infrastructure will improve the accessibility of telehealth capabilities and extend the interoperability of the digital reality platform allowing for a diverse digital health care ecosystem.

Acknowledgments
This review was supported by the Science Foundation Ireland under grants (13/RC/2073) and (13/RC/2094). The authors would like to thank the university librarian, Rosie Dunne, for providing feedback on the search strategy and helping in finding the appropriate databases to conduct the search in.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 29 KB - biomedeng_v8i1e42709_app1.docx ]

Multimedia Appendix 2
Augmented, virtual, and mixed reality hardware devices.
[DOCX File, 22 KB - biomedeng_v8i1e42709_app2.docx ]

Multimedia Appendix 3
Other relative hardware devices.
[DOCX File, 22 KB - biomedeng_v8i1e42709_app3.docx ]

Multimedia Appendix 4
List of software and source platforms.
[DOCX File, 27 KB - biomedeng_v8i1e42709_app4.docx ]

References
3. Telehealth and Telemedicine. American Academy of Family Physicians. URL: https://www.aafp.org/about/policies/all/telehealth-telemedicine.html [accessed 2021-12-01]


https://biomedeng.jmir.org/2023/1/e42709


Abbreviations

AR: augmented reality
HMD: head-mounted display
H-TIME: Haptic Enable Tele-Immersion Musculoskeletal Examination
ICT: information and communication technology
LCD: liquid crystal display
LED: light emitting diode
MR: mixed reality
POCUS: point-of-focus ultrasonography
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
REBOA: resuscitative endovascular balloon occlusion of the aorta
ReCoVR: Realizing Collaborative virtual reality for well-being and self-healing
STAR: System for Telementoring with Augmented Reality
USG: ultrasonography
Development and Testing of a Data Capture Device for Use With Clinical Incentive Spirometers: Testing and Usability Study

Michael L Burns1, MD, PhD; Anik Sinha1, MSc; Alexander Hoffmann1, BSc; Zewen Wu1, MSc; Tomas Medina Inchauste1, MSc; Aaron Retsky1, BSc; David Chesney2, PhD; Sachin Kheterpal1, MBA, MD; Nirav Shah1, MBA, MD

1Department of Anesthesiology, University of Michigan Health System, Ann Arbor, MI, United States
2Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Michael L Burns, MD, PhD
Department of Anesthesiology
University of Michigan Health System
1500 E Medical Center Drive
Ann Arbor, MI, 48109
United States
Phone: 1 734 936 4171
Email: milburns@med.umich.edu

Abstract

Background: The incentive spirometer is a basic and common medical device from which electronic health care data cannot be directly collected. As a result, despite numerous studies investigating clinical use, there remains little consensus on optimal device use and sparse evidence supporting its intended benefits such as prevention of postoperative respiratory complications.

Objective: The aim of the study is to develop and test an add-on hardware device for data capture of the incentive spirometer.

Methods: An add-on device was designed, built, and tested using reflective optical sensors to identify the real-time location of the volume piston and flow bobbin of a common incentive spirometer. Investigators manually tested sensor level accuracies and triggering range calibrations using a digital flowmeter. A valid breath classification algorithm was created and tested to determine valid from invalid breath attempts. To assess real-time use, a video game was developed using the incentive spirometer and add-on device as a controller using the Apple iPad.

Results: In user testing, sensor locations were captured at an accuracy of 99% (SD 1.4%) for volume and 100% accuracy for flow. Median and average volumes were within 7.5% (SD 6%) of target volume sensor levels, and maximum sensor triggering values seldom exceeded intended sensor levels, showing a good correlation to placement on 2 similar but distinct incentive spirometer designs. The breath classification algorithm displayed a 100% sensitivity and a 99% specificity on user testing, and the device operated as a video game controller in real time without noticeable interference or delay.

Conclusions: An effective and reusable add-on device for the incentive spirometer was created to allow the collection of previously inaccessible incentive spirometer data and demonstrate Internet-of-Things use on a common hospital device. This design showed high sensor accuracies and the ability to use data in real-time applications, showing promise in the ability to capture currently inaccessible clinical data. Further use of this device could facilitate improved research into the incentive spirometer to improve adoption, incentivize adherence, and investigate the clinical effectiveness to help guide clinical care.

(JMIR Biomed Eng 2023;8:e46653) doi:10.2196/46653

KEYWORDS

incentive; spirometry; Internet-of-Things; electronic health records; web-based intervention; medical device; medical tool; data collection; spirometry data; incentive spirometer; data analysis; algorithm; effectiveness

Introduction

Pulmonary complications after major surgery, including pneumonia, atelectasis, respiratory failure, prolonged supplemental oxygen requirements, and reintubation, are common, expensive, and harmful to patients [1]. Studies estimated these complications in the range of 2%–39% [2,3], though atelectasis alone has been found to affect up to 92% of...
postsurgical patients [4]. Originating in the 1970s, the incentive spirometer was designed to mimic the physiology of a sigh or a yawn—a slow voluminous inhalation [5,6], and this basic medical device is often used in postoperative care to aid lung expansion to prevent or reduce respiratory complications.

To complicate the matter, correct use and device adherence is low among patients. Guidelines on how to properly use the incentive spirometer device outline best practices [7,8], but over 26% of patients fail to use their incentive spirometer correctly, and over 38% deny ever using their device in their postoperative care [9], highlighting the need for more evidence-based recommendations. There is sparse evidence for the use of the incentive spirometer device [10] for postoperative pulmonary complication prevention [11] with only a few studies demonstrating clinical effectiveness when used properly [12,13]. As a result of the lack of high-level evidence, some clinical practice guidelines do not support its routine postoperative use [14]. Additionally, there remains disagreement as to the most effective way to use the device, as studies have been unable to demonstrate its superiority over other techniques such as deep breathing techniques, directed coughing, early mobilization, and optimal analgesia [15]. Uncertainty around the effective spirometry use is partially due to the scarcity of spirometer compliance data [16]. Compliance measurements, made through self-reporting and staff observation, are difficult to obtain, and when captured, they have demonstrated low patient adherence to the incentive spirometer device [12]. Though data remain elusive, 86% of health care providers believe patient adherence is poor, and 95.4% believe it should be improved [17], demonstrating the perceived use of incentive spirometry. The first step in determining the optimal use of the incentive spirometer is to improve data collection. Automated capture of spirometry data may improve the quality of research studies and ultimately determine the incentive spirometer’s use in improving lung function and minimizing postoperative complications.

Incentive spirometers are simple plastic meters to measure inhalational breath flow and volume; they lack the ability to record data. Digital flowmeters can be used to replicate the incentive spirometer [18]. While an option for improving data capture, the digital flowmeter can be clinically infeasible in most practices due to its complexity and cost. Collecting data directly from incentive spirometers falls into a technology category called the “Internet-of-Things” [19]. There have been great advances in the miniaturization of computing devices and the evolution of the “Internet-of-Things” into mainstream health devices [20,21]. As a result, wearable technologies and web-based platforms are capturing more clinical data now than ever before [22,23]. These technologies give clinicians and researchers access to otherwise inaccessible patient data and the ability to investigate new data interactions, such as enabling patients to exercise with wearable device hardware (eg, chest monitors and watches) and incorporate and communicating these data with patient physiologic and movement data [24,25]. The goal was to use this technology for the incentive spirometer.

We hypothesize an add-on device can accurately measure flow and volume data from a common incentive spirometer. This paper describes the creation and testing of an incentive spirometer add-on device to measure flow and volume data. An Internet-of-Things approach was adopted to enable this device to work with an existing incentive spirometer to capture physiological data and communicate externally on a closed private internet connection using hospital Wi-Fi and Bluetooth technologies. Captured data were tested for use by developing a classification algorithm to determine valid from invalid breaths.

**Methods**

**Ethics Approval**

Institutional review board approval was obtained for this prospective exploratory study on April 15, 2021 (HUM00196543, University of Michigan, Ann Arbor). Informed consent was waived as participants were limited to authors within this study, and interventions for this study are limited to existing approved uses of the incentive spirometer.

**Add-On Device Creation**

An add-on device was designed and created for use with the Hudson RCI Voldyne 5000 Volumetric Exerciser (Teleflex Medical) incentive spirometer (Figures 1-3 and Multimedia Appendix 1) and composed of photoelectric reflective infrared optical sensors (Xingyheng) positioned lateral to the spirometer volume and flow columns. Ten sensors were placed along the spirometer volume column with each sensor along a 500 mL marking on the incentive spirometer device (500-5000 mL). Three sensors were placed along the flow column corresponding to the middle of flow spirometer markings (“best,” “better,” and “good”). Sensors were connected to a microcontroller by a solderless breadboard, using an ESP32 development board (Dorhea), a low-cost low-power system-on-a-chip microcontroller with integrated Wi-Fi and dual-mode Bluetooth wireless communication capability. Components were soldered on a breadboard with breadboard jumper wires and resistors connecting components. Interactions between the microcontroller and the sensors were coded using C++ (Bell Laboratories of American Telephone and Telegraph). Components were encased in a 3D-printed base situated beneath the incentive spirometer with a layer of plexiglass fixed in the middle.

Data were sent directly to an iOS application running on an iPad Pro (12.9” and 10.2”, fifth generation, Apple) through ESP32 Bluetooth as well as stored on the add-on device in a 32 gigabyte microsecure digital card (Kootion) and card reader module (HiLetGo), a clock module (Melife) time-stamped data. Further interconnection was made possible using an analog digital multiplexer breakout board (Xie QianJin) and included a 3.7-V 2400-mAh rechargeable lithium battery (Akzytue) and charging module (MakerFocus) to power the device. An alternative design for the add-on base can be found in Multimedia Appendix 2.
Figure 1. The incentive spirometer piston positions were read using reflective infrared optical sensors (1) and relayed to an ESP32 in the base of the add-on device (2). The device both stored the data internally and used Bluetooth technology to transmit spirometer data to an Apple iPad (3). Data were then transmitted from the iPad to dedicated servers for further data processing and storage (4).

Figure 2. Ten photoelectric reflective infrared optical sensors were placed along the spirometer volume column with each sensor along a 500 mL marking on the incentive spirometer device (500-5000 mL); 3 sensors placed along the flow column. An on/off switch is featured on the front of the device. An ESP32 development board, rechargeable lithium battery, and real-time clock module are labeled. Connections between components were made with breadboard jumper wires and resistors, and all components are encased in a 3D-printed base with a layer of plexiglass overtop. A microsecure digital card reader module is in the back of the base (not shown).
Software and Video Game Creation

An iOS video game and analytics application was built using Unity and C# (Microsoft). A web server receives the data and handles application programming interface (API) requests written in Java (Sun Microsystems). The ESP32 reads the sensor data using custom C++ code and sends processed data to an Apache web server via Wi-Fi through an iPad device connected via Bluetooth (Figure 1). Python (Python Software Foundation) applications were exposed by the web server, and downstream applications were networked with the server through an API to allow data collection and further use. A web server receives the data and handles API requests. A video game was developed specifically for use with the incentive spirometer and add-on device serving as a controller, designed for use with an Apple iPad Pro.

Add-On Device Sensor Testing

Two investigators (AS and MLB) tested the add-on device without crossover using the Hudson RCI Voldyne 5000 Volumetric Exerciser incentive spirometer. In these tests, 5 user breaths were attempted at each of the 10 volume sensor positions (100-5000 mL) for a total of 50 breaths per user. The volume goal was to get the top of the volume piston to the desired volume marking, while the flow goal was to get the top of the flow bobbin to the middle of the desired flow marking. Flow readings were also tested with every volume test with an additional 5 flow tests at each of the 3 flow sensor positions. To attain the desired flow or volume piston levels, users were allowed to breathe through or tilt the incentive spirometer device.

A single investigator (MLB) tested flow and volume sensor ranges with the incentive spirometer connected to the Puritan-Bennet PTS-2000 Ventilator Analyzer Tester (Mallinckrodt) digital flowmeter. In this testing, the investigator attempted increasing volume and flow values to identify sensor ranges. Volume measurements from the flowmeter were corrected using the body temperature, pressure, water vapor saturated method (correcting for body temperature [37 °C], ambient pressure, and gas saturated with water vapor). The flowmeter volume was calculated to be 183.75 mL by calibrating the flowmeter readings precisely at the 1000 mL display level of the Voldyne 5000, averaged across 10 breaths. This volume was subtracted from each raw volume reading from the flowmeter. Flowmeter testing was attempted with a minimum of 25 breaths around each flow and volume sensor level, independently, and the volume achieved from the flowmeter and the sensor level attained from the iPad were recorded. Due to limitations continually achieving breath volumes above 3250 mL (sensor level 7), additional testing using the 2500 mL volume Hudson RCI Voldyne 2500 Volumetric Exerciser (Teleflex Medical) incentive spirometer was conducted. The add-on device was created with each sensor aligning to the 500 mL markings of the Voldyne 5000. When using the Voldyne 2500, the sensors were near but incompletely aligned to the 250-mL markings.

Breath Algorithm Development and Testing

While sensors in the add-on device can detect where a piston is located, they cannot determine if a breath was conducted to achieve the sensor level. It is important to be able to accurately identify when a breath moves the incentive spirometer piston.
and bobbin as opposed to tilting the incentive spirometer causing them to fall to the desired level. To aid in distinguishing a breath attempt, an algorithm was developed in Python (version 3.1) to process and classify each user's breath using flow and volume data over time. This algorithm reads spirometer log data into a pandas data frame and parses data by identifying the start and end of breaths. Breath start was identified by a zero flow value that precedes a positive flow value, while breath end is identified by 2 consecutive zero flow values. Once identified, each breath is classified as “valid” or “invalid.” A valid breath must have the following volume criteria: volume starts at 0, increases within 1.5 seconds of breath start, >0 throughout the breath, does not decrease while there is positive flow, and the length of the breath is between 2 and 15 seconds. Breaths that do not meet these requirements were deemed invalid.

To test this algorithm, separate from device testing, 2 investigators (AS and MLB) each attempted to create 5 valid and 5 invalid breaths at each volume sensor level. These breath data were processed through the classification algorithm and evaluated. Invalid breaths were created by manually tipping the incentive spirometers or starting breaths at a starting position >0. All data validation was retrospectively validated using expert opinion from one of 2 users (AS and MLB).

Statistical Analysis

Summary and descriptive statistics were calculated for collected data using basic statistical techniques to assess models created for breath analysis including accuracy (as defined by user identification as gold standard), as well as mean, median, and SD for volume and flow measurements in sensor range testing.

Results

Add-On Device Sensor Testing

An add-on incentive spirometer device was created to independently measure real-time spirometer flow and volume piston positions (Figures 1-3). From 2 investigators, 99 of 100 volume readings were measured at the correct corresponding volume sensor (accuracy 99%, SD 1.4%). Only a single volume reading was inaccurate, failing once to capture at the lowest volume sensor (500 mL). All 130 flow bobbin readings corresponded to the correct flow sensor (100% accuracy).

Volume and flow sensor ranges were determined using a digital flowmeter in line with the 5000 and 2500 mL incentive spirometers independently (Figure 4 and Tables 1 and 2). Add-on sensors were designed with placement at every 500 mL marking on the Voldyne 5000 incentive spirometer. These sensor placements were not adjusted for the 2500 mL incentive spirometer testing. Investigator breath limitations were reached at volume levels approaching 4000 mL. Between the 2 devices tested, median and average volumes were within 7.5% (SD 6%) of target volume sensor levels. Maximum triggering values seldom exceeded the intended sensor level, showing a good correlation to placement. Flow levels corresponding to “best,” “better,” and “good” levels differed significantly between devices with the 5000 mL flow values more than double the 2500-mL incentive spirometer at each level.

Figure 4. User testing of add-on device connected to a digital flowmeter for both the Voldyne 5000 (A) and the Voldyne 2500 (B) incentive spirometers. Circles represent individual breath attempts with digital flowmeter volume readings (L, x-axis) and add-on device sensor iPad readings (y-axis). Sensor levels were designed to be placed at 500 mL levels of the Voldyne 5000 incentive spirometer and closely aligned to 250 mL markings of the Voldyne 2500.
Table 1. Volume sensor ranges.a

<table>
<thead>
<tr>
<th>Number</th>
<th>Breath, n</th>
<th>Volume (L)</th>
<th>Mean (SD)</th>
<th>Median (IQR)</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voldyne 5000 volume sensor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>43</td>
<td>0.61 (0.10)</td>
<td>0.62 (0.54-0.67)</td>
<td>0.43</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>71</td>
<td>0.98 (0.20)</td>
<td>0.99 (0.81-1.13)</td>
<td>0.63</td>
<td>1.55</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>1.55 (0.18)</td>
<td>1.54 (1.40-1.68)</td>
<td>1.27</td>
<td>1.92</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>2.20 (0.12)</td>
<td>2.17 (2.11-2.28)</td>
<td>1.99</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>88</td>
<td>2.59 (0.20)</td>
<td>2.59 (2.42-2.76)</td>
<td>2.21</td>
<td>2.98</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>37</td>
<td>3.09 (0.13)</td>
<td>3.10 (3.02-3.16)</td>
<td>2.79</td>
<td>3.38</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>48</td>
<td>3.59 (0.35)</td>
<td>3.53 (3.29-3.96)</td>
<td>3.00</td>
<td>4.18</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>4.29 (N/A)</td>
<td>4.29 (N/A)</td>
<td>4.29</td>
<td>4.29</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>N/A (N/A)</td>
<td>N/A (N/A)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>N/A (N/A)</td>
<td>N/A (N/A)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Voldyne 2500 volume sensor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>0.21 (0.04)</td>
<td>0.22 (0.18-0.23)</td>
<td>0.13</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>0.41 (0.13)</td>
<td>0.43 (0.29-0.48)</td>
<td>0.22</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>0.73 (0.11)</td>
<td>0.72 (0.64-0.82)</td>
<td>0.57</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>1.03 (0.09)</td>
<td>1.02 (0.99-1.09)</td>
<td>0.87</td>
<td>1.20</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>1.26 (0.10)</td>
<td>1.27 (1.18-1.34)</td>
<td>1.09</td>
<td>1.43</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>1.60 (0.11)</td>
<td>1.61 (1.53-1.67)</td>
<td>1.38</td>
<td>1.78</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>1.89 (0.11)</td>
<td>1.91 (1.81-1.96)</td>
<td>1.71</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>2.21 (0.11)</td>
<td>2.24 (2.13-2.28)</td>
<td>1.99</td>
<td>2.38</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>2.47 (0.11)</td>
<td>2.49 (2.40-2.55)</td>
<td>2.24</td>
<td>2.63</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>37</td>
<td>2.75 (0.14)</td>
<td>2.74 (2.64-2.85)</td>
<td>2.52</td>
<td>3.09</td>
<td></td>
</tr>
</tbody>
</table>

aResults from investigations of increasing volume sensor readings from the Voldyne 2500 and Voldyne 5000 incentive spirometers. The number of breaths (n) at each sensor level (1-10) and the average, median (IQR), minimum, and maximum readings were obtained from a digital flowmeter in liters (L).

bN/A: not applicable.

Table 2. Flow sensor ranges.a

<table>
<thead>
<tr>
<th>Breath, n</th>
<th>Flow (L/minute)</th>
<th>Mean (SD)</th>
<th>Median (IQR)</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voldyne 5000 flow sensor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>30</td>
<td>26.36 (5.56)</td>
<td>26.04 (22.51-27.76)</td>
<td>19.21</td>
<td>40.21</td>
</tr>
<tr>
<td>Better</td>
<td>30</td>
<td>44.31 (5.49)</td>
<td>43.56 (40.44-48.25)</td>
<td>36.10</td>
<td>56.69</td>
</tr>
<tr>
<td>Good</td>
<td>31</td>
<td>57.75 (7.74)</td>
<td>58.24 (53.99-61.27)</td>
<td>39.36</td>
<td>71.38</td>
</tr>
<tr>
<td>Voldyne 2500 flow sensor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>30</td>
<td>12.24 (2.36)</td>
<td>12.31 (10.69-13.12)</td>
<td>8.43</td>
<td>19.63</td>
</tr>
<tr>
<td>Better</td>
<td>30</td>
<td>18.70 (1.92)</td>
<td>19.08 (17.45-19.56)</td>
<td>15.07</td>
<td>22.60</td>
</tr>
<tr>
<td>Good</td>
<td>31</td>
<td>28.45 (3.34)</td>
<td>28.13 (25.93-29.96)</td>
<td>22.61</td>
<td>38.30</td>
</tr>
</tbody>
</table>

aResults from investigations of increasing flow sensor readings from the Voldyne 2500 and Voldyne 5000 incentive spirometers. The number of breaths (n) at each sensor level (“best,” “better,” and “good”) and the average, median (IQR), minimum, and maximum readings were obtained from a digital flowmeter in liters per minute (L/minute).
Breath Algorithm Development and Testing

After data collection, an algorithm was developed to classify breath data to determine when the spirometer was actively used. This algorithm works by classifying each individual breath using criteria to determine their validity as described in the Methods section. Independent of sensor testing, 2 members of the study team attempted 5 valid and 5 invalid breaths at each of ten 500-mL volume levels (500-5000 mL). Investigator limitations in achievable volumes resulted in a total of 65 valid breaths and 100 invalid breaths. The breath classification algorithm resulted in a 100% sensitivity and a 99% specificity for the classification of “valid” versus “invalid” user breaths. A single valid reading was inappropriately classified by the algorithm, occurring at the 500 mL volume sensor level. Example breath algorithm results are shown in Figure 5.

To investigate the downstream applicability of breath data, a video game was developed using Unity software, Apple iOS, with a single final application installed on Apple iPad devices (Figure 6). The game was a Kirby (Nintendo)-based side scroller, where the character would approach an obstacle and traverse the obstacle after a successful breath. This game was developed specifically for use with the incentive spirometry device with gameplay centered around proper use [7,8]. The incentive spirometer with an add-on device was successfully used to control the created game. A video showing the game played in real time with an incentive spirometer device can be found in Multimedia Appendix 3.

Figure 5. Example results from the breath algorithm used to classify “valid” from “invalid” breath attempts. The y-axis represents sensor levels for flow and volume readings in the incentive spirometer device: flow (0-3; corresponding to none: 0; “good”: 1; “better”: 2; and “best”: 3); volume (0-10, representing each 500 mL increment from 0 to 5000 mL). The x-axis represents the time from breath start in seconds.

Figure 6. A video game was specifically designed for use with the incentive spirometer and add-on device based on the Kirby Nintendo character. A sprite sheet for the character development within the video game is shown on the left with a screenshot of the video game shown on the right. Multimedia Appendix 3 shows the game being controlled using an incentive spirometer.
Discussion

Principle Findings

In this study, an add-on device was created to allow the capture of incentive spirometry data with high accuracy (99%, SD 1.4% volume and 100% flow when tested at incentive spirometer level markings). Maximum triggering values were rarely exceeding intended sensor levels, proving excellent differentiation of levels in the design. Furthermore, while sensor placements were designed around the Voldyne 5000 incentive spirometer, the similarly shaped Voldyne 2500 performed well in testing without sensor modifications, suggesting use of the add-on device to similar incentive spirometers without significant redesign. While range testing was not intended to study volume sensor accuracy, between the 2 spirometers tested, median and average volumes were within 7.5% (SD 6%) of target volume levels, with the worst individual readings at lower volumes (250-500 mL sensors) with readings maximally 40 to 120 mL off target, respectively.

To differentiate quality breath attempts from errors, a classification breath algorithm was developed. The device allowed identification of start and end of breath attempts, valid from invalid breaths assessment, maximum flow, maximum volume, and volume or time ascents and descent calculations. The breath classification algorithm used device data to discern valid versus invalid breath attempts, showing 100% sensitivity and 99% specificity. Identifying valid breath data from noise is critically important for downstream applications and, while the classification algorithm yielded results, further efforts for improvement could be made using additional rule-based systems or machine learning algorithms.

This system is intended to enable clinical providers access to previously inaccessible spirometry data to improve spirometer instruction and use protocols, study patient compliance, and incentivize use. Using an add-on device similar to what was created in this study would increase the granularity of spirometer compliance data and could be used to provide insight into proper incentive spirometer use. Additionally, the add-on device can allow focused interventions to improve adherence. Reminder notifications alone have been shown to improve incentive spirometer use. In one study, an add-on use-tracking device was equipped with a bell that sounded for up to 2 minutes every hour as a reminder for the patient to use their incentive spirometer [12]. This study demonstrated that patients using the reminder device had a greater number of mean daily inspiratory breaths and a percentage of recorded hours with an inspiratory breath event. More importantly, patients with the reminder displayed significantly lower mean atelectasis severity scores measured by chest radiography, reduced median postoperative and intensive care unit length of stay, and had a lower mortality rate at 6 months. These findings support postoperative incentive spirometer use and show effectiveness of a simple intervention to improve incentive spirometer adherence.

Gamification

To demonstrate the real-time use of incentive spirometry data, iPad video game was created to be controlled by the add-on incentive spirometer device. In testing, the game showed no appreciable lag and continued connectivity during use, proving electronic spirometer data collected by the add-on device to be capable of real-time gamification applications. Gamification of medical interventions is an exciting concept for improving medical care adherence. Breathing games for the incentive spirometer is a familiar idea, with one group brainstorming a suite of games for asthmatics focusing on breathing metaphors as incentives for spirometer use [26], while others developed video games to incentivize breathing exercises and peak expiratory flow using digital flowmeters [18,27]. There exists an abandoned patent around the use of the incentive spirometer as a game controller [28] and an active patent around use of the flowmeter in video games [29], further supporting the popularity of the idea. The device in this study was created leveraging recent technologies and focusing design on clinical care use. Using the add-on device, as opposed to a digital flowmeter, maintains the current use of incentive spirometers in medical settings to allow native data capture. Potential reuse of the add-on device limits additional costs such as those incurred using digital flowmeters.

Limitations

Limitations to the add-on device design include contamination risks, costs, and technical and workflow implementation constraints. First, while completely enclosed, the add-on device is designed to be reused and carries the risk of infection—an especially important consideration in a respiratory pandemic such as COVID-19. To improve sterility, the device was enclosed in plastic, and sensors were placed behind the incentive spirometer, removing the need to expose the base to breaths from sensors placed below. Improvements can be made to close remaining gaps in the plastic encasing to further enclose the device and allow cleaning like an iPad or PlayStation controller, commonly used in the hospital settings. Second, routine incentive spirometer postoperative care has been estimated to carry a US $107.36 cost per patient above material cost of the spirometer device, which totals US $1.04 billion in total US annual costs [30]. This is a significant cost for a device with sparse evidence around use and poor patient compliance. While incentive spirometer devices are not reusable from patient to patient, the add-on device was designed to be reusable, lowering its effective cost. Overall, the add-on device carries a material price of approximately US $150 per unit (Multimedia Appendix 4), which could be reduced by bulk purchasing and further investigation into alternative individual components. While add-on device cost is an addition to the already significant price, identifying compliance and improving adherence will facilitate improved use and function of the incentive spirometer. Further studies of the incentive spirometer are required to investigate the prevention of postoperative breathing complications and their associated health care costs. These studies are dependent upon accurate compliance data and would benefit from the capabilities of the add-on device. Third, there exist integration and maintenance requirements of the add-on device, and it may be feasible for use only in hospital systems with existing technical support structures. The device was designed to minimize technical requirements, but more investigation is required. Future studies are required to trial the add-on device.
in clinical settings and test for improved adherence using strategies such as gamification compared to traditional incentive spirometry.

Conclusions
Incentive spirometers are routinely used in hospital settings, specifically in postoperative clinical care, but recommendations for proper routine use lack thorough investigation due to a general lack of data on device use. Creating a low-cost, effective, and reusable add-on device for the incentive spirometer allows native collection of previously inaccessible incentive spirometer compliance data. These data can facilitate research into incentive spirometer use to guide clinical care, incentivize adherence, and draw conclusions about the clinical effectiveness of the incentive spirometer.

Acknowledgments
The authors gratefully acknowledge the valuable contribution of Mikele Garrett and dedicate this paper in his memory. Mikele was an incredible person and teammate who passed away prior to the completion of this paper. He was critical in device development, overseeing and creating the add-on device designs. The authors also acknowledge the Adult Clinical Research Group at the University of Michigan for the valuable contribution to protocol development and final paper review; the University of Michigan Hacks with Friends competition for initial idea development; and the University of Michigan Software Development for Accessibility course EECS 495 and its students for video game ideas and creation, Amanda Bahamonde for her assistance with preparing this paper for submission, Ruth Cassidy for statistical support, and Kent Miller for access to the digital flowmeter used in device testing. Partial funding can be attributed to the Department of Anesthesiology, University of Michigan Medical School, the Society for Technology in Anesthesia 2021 Neurowave Research Grant, and the National Institute for General Medical Sciences of the National Institutes of Health (MLB; grant T32GM103730).

Data Availability
Deidentified data were collected from spirometer use by the study team. Data were collected using the incentive spirometer, add-on device, and digital flowmeter. Data sets generated and analyzed within this study are available from the corresponding author upon reasonable request.

Conflicts of Interest
This work has been declared through the University of Michigan Innovation Partnerships.

Multimedia Appendix 1
Add-on device design and schematics.
[DOCX File, 294 KB - biomedeng_v81e46653_app1.docx ]

Multimedia Appendix 2
Alternative design schematic for add-on device.
[DOCX File, 436 KB - biomedeng_v81e46653_app2.docx ]

Multimedia Appendix 3
Video displaying an investigator (TM) using the incentive spirometer with the add-on device to control the created video game.
[MP4 File (MP4 Video), 7764 KB - biomedeng_v81e46653_app3.mp4 ]

Multimedia Appendix 4
Add-on device cost.
[DOCX File, 13 KB - biomedeng_v81e46653_app4.docx ]

References


Abbreviations

API: application programming interface