

# Metrics for Assessing the Reliability of a Telemedicine Remote Monitoring System

Neil Charness, PhD,<sup>1</sup> Mark Fox, PhD,<sup>1</sup> Amy Papadopoulos, DSc,<sup>2</sup> and Cindy Crump, MA<sup>2</sup>

<sup>1</sup>Department of Psychology, Florida State University, Tallahassee, Florida.

<sup>2</sup>AFrame Digital Inc., Reston, Virginia.

## Abstract

**Objective:** The goal of this study was to assess using new metrics the reliability of a real-time health monitoring system in homes of older adults. **Materials and Methods:** The “MobileCare Monitor” system was installed into the homes of nine older adults >75 years of age for a 2-week period. The system consisted of a wireless wristwatch-based monitoring system containing sensors for location, temperature, and impacts and a “panic” button that was connected through a mesh network to third-party wireless devices (blood pressure cuff, pulse oximeter, weight scale, and a survey-administering device). To assess system reliability, daily phone calls instructed participants to conduct system tests and reminded them to fill out surveys and daily diaries. Phone reports and participant diary entries were checked against data received at a secure server. **Results:** Reliability metrics assessed overall system reliability, data concurrence, study effectiveness, and system usability. Except for the pulse oximeter, system reliability metrics varied between 73% and 92%. Data concurrence for proximal and distal readings exceeded 88%. System usability following the pulse oximeter firmware update varied between 82% and 97%. An estimate of watch-wearing adherence within the home was quite high, about 80%, although given the inability to assess watch-wearing when a participant left the house, adherence likely exceeded the 10h/day requested time. In total, 3,436 of 3,906 potential measurements were obtained, indicating a study effectiveness of 88%. **Conclusions:** The system was quite effective in providing accurate remote health data. The different system reliability measures identify important error sources in remote monitoring systems.

**Key words:** home health monitoring, e-health, telehealth, telemedicine reliability measures identify

## Introduction

Data integrity is critical to providers who interpret remotely collected data streams.<sup>1</sup> False-positives by equipment used in pediatric intensive care units have been shown to exceed 68% and together with inadvertent triggering can reach 90%.<sup>2,3</sup> Even more serious are false-negatives, failing to detect

a critical event.<sup>4</sup> Assessing the reliability of a system is an important first step in determining its potential effectiveness.

There appear to be few standards for assessing reliability of remote monitoring systems. For passive monitoring systems (e.g., implantable cardiac devices<sup>5</sup>), reliability is usually assessed by checking received data from sensor systems (such as daily messages) against expected messages (what we call *data concurrence*), or alerts from intelligent system are validated against later hospital-assessed examination of device status (what we call *system reliability*). Determining reliability is even more difficult for hybrid systems with both passive (e.g., worn or automatically transmitted) and active (e.g., patient-activated devices, such as blood pressure [BP] cuffs, glucose meters, weight scales) components. For example, a study of elderly patients required patients to monitor blood glucose levels four times daily, but patients only averaged twice-daily compliance as assessed by a system that transmitted use data wirelessly.<sup>6</sup> (See the recent review of wearable remote monitoring systems of Patel et al.<sup>7</sup>)

Hence, an interrelated factor in system effectiveness is usability, the extent to which people can quickly learn to operate the system reliably and safely. Failures in human components can be even more harmful than failures in machine components.<sup>8,9</sup> In brief, system reliability measures need to take the human factor into consideration.<sup>10</sup> For hybrid systems, developing reliability metrics that can differentiate failures associated with human components from automated system components could lead to better training of system users.

We report on a real-time monitoring and alerting system, the “MobileCare Monitor,” that was deployed into nine people’s homes for 2 weeks to assess reliability and usability of some of its components, taking both instrument and human failure sources into account. The system combined the wireless wristwatch-based monitoring system from AFrame Digital (Reston, VA) ([www.aframedigital.com](http://www.aframedigital.com)),<sup>11</sup> containing sensors to assess location within a home (using a mesh network), temperature, impacts (e.g., falls), and a manual “panic” button function, with third-party wireless devices: a BP cuff, a weight scale, a pulse oximeter, and a WiFi survey-administering device (Chumby Industries, Inc., San Diego, CA), containing a touch-sensitive screen (finger or stylus activation) and programmed with eight daily questions (SF-8) derived from the SF-36 health questionnaire.

## Materials and Methods

### PARTICIPANTS

Following an Institutional Review Board-approved procedure, individuals in Leon County, Florida were contacted through

advertising, from referrals, and from a re-contact permission database. Inclusion criteria were as follows: being 75 years of age or older, possession of a home broadband connection, availability for 2 weeks, and being free of dementia (determined by the short portable mental status questionnaire and the Wechsler memory scale). Nine participants >75 years of age were enrolled, three at a time, between February 11 and May 30, 2010. They were paid \$25/day for the 14-day study.

**EQUIPMENT**

The installed system contained components from AFrame including their watch device, third-party Bluetooth® (Bluetooth SIG, Kirkland, WA) protocol devices connected via a Bluetooth gateway to AFrame’s system, and the WiFi-connected survey device (Chumby). Data from the system were conveyed in encrypted format over the Internet to secure servers in Virginia. *Table 1* shows the list of components, and *Figure 1* shows their interconnections.

In cases of home power failure (one did occur during the study), AFrame system devices automatically reset on power resumption. Passive sensor components in the system included watch temperature readings, watch location information, watch battery status, and watch impact events (accelerometer-based detection). All other

measurements are active, requiring participant initiation (e.g., pulse, BP, oxygen saturation, etc.), although they are transmitted automatically through the system. Bluetooth protocol was used to connect third-party devices via the Bluegiga Bluetooth access point to the Zigbee-based mesh network.

**PROCEDURE**

Installation was preceded by having the participant read and sign an informed consent form. Installation typically took between 1 and 2 h using AFrame’s procedure.

At installation, the participant was provided with a diary document for manually recording physiological readings taken at roughly the same time daily for a 2-week period and given instruction on how to use the equipment. Participants were provided with cell phone numbers for technical support. On Day 1 esthetics questionnaires for the Chumby and watch were administered.

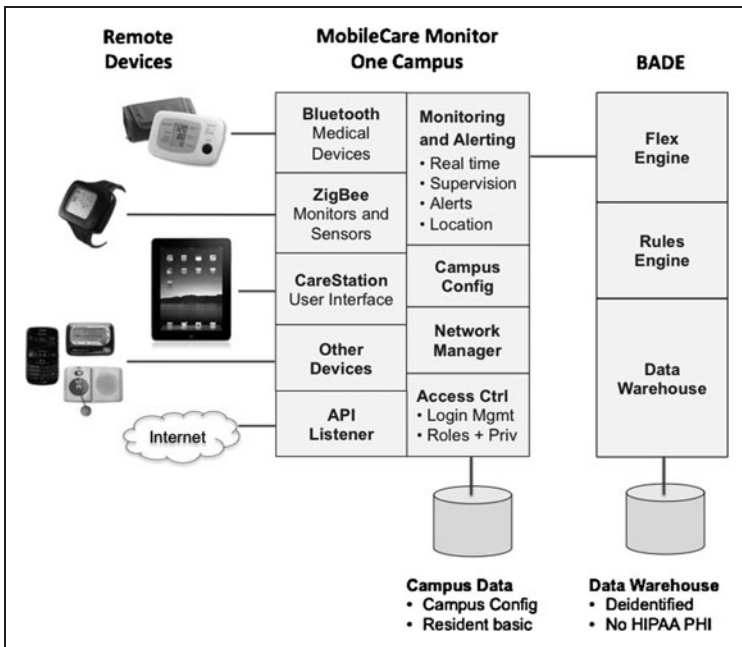
The participants were asked to wear the AFrame watch at least 10 h/day, between 8 a.m. and 8 p.m. In addition, participants were phoned daily (usually in the early evening or morning) to test the watch panic button function, to check their BP, pulse oximetry, and weight, and to be reminded to take the Chumby health questionnaire. During the call, participants were reminded to fill out the daily

**Table 1. List of Equipment Installed in Homes**

DEVICE	MANUFACTURER	MODEL/NAME	FDA-CLEARED
AFrame Digital myPHD watch	AFrame	myPHD Wearable Monitor, Grape	Yes
AFrame Digital myPHD travel charger	AFrame	myPHD travel charger	Yes
Xbee-Pro ZB wall router	Digi	XR-Z14-CW1P1	Yes
ConnectPort X4	Digi	X4-Z11-E-A	Yes
Bluetooth access point (gateway)	Bluegiga	3201	NA
Bluetooth weight scale	A&D	UC-321PBT	Yes
Bluetooth blood pressure monitor	A&D	UA-767PBT	Yes
Bluetooth fingertip pulse oximeter	Nonin	9560 OnyxII <sup>a</sup>	Yes
Chumby classic touchscreen Web interface	Chumby	CHY-A01-A	NA
"MOBILECARE MONITORING" SYSTEM SOFTWARE AND FIRMWARE		VERSION	
AFrame CareStation (secure Web application)		SW version: MCM 3.3.4 2010-01-07	
AFrame BADE database		SW version: BADE 2.1	
AFrame digital myPHD watch		FW version: 2.28	
AFrame digital myPHD travel charger		XA03504 revision 1	
Xbee-Pro ZB wall router		FW version: 2241	
ConnectPort X4		FW version: 2141	
ConnectPort X4		SW version: PANDA 2.05	

<sup>a</sup>The Nonin 9560 OnyxII wireless devices were replaced by the manufacturer partway through the study because of issues with excessive battery consumption. Five study participants used the older device with battery issues, and four used the replacement devices with updated firmware.

FDA, Food and Drug Administration; FW, firmware; NA, not applicable; SW, software.



**Fig. 1.** Schematic of the telehealth system. API, application program interface; HIPAA PHI, Health Insurance Portability and Accountability Act, protected health information.

notebook, which included comfort ratings for the watch as well as the daily weight, heart rate, oxygen saturation level, and BP readings. CareStation readings were monitored during the phone call from a remote computer using the Web browser interface. Experimenter notes were taken during calls.

At the end of each 2-week trial, Florida State University personnel collected the equipment and daily journals from the participant after administering both the esthetics questionnaire and the comfort rating

scale. At the completion of testing, participants were compensated \$25/day for the 14-day study.

At study completion, AFrame personnel downloaded all data received automatically by the AFrame server and forwarded data to Florida State University for comparison with the daily logs provided by the study participant and notes recorded by Florida State University personnel; the latter information was blinded from AFrame personnel. The data comparisons for values in which data was gathered both manually and by the AFrame system are shown in *Table 2*. Readings for the Nonin Medical, Inc. (Plymouth, MN) pulse oximeter device are shown in two categories: one with the original device (first five installations) and one that used the replacement device (final four installations).

**Results**

**THIRD-PARTY PHYSIOLOGICAL MEASUREMENT DEVICES**

One hundred twenty-six data points were possible for BP, heart rate, blood oxygenation, and weight (nine participants for 14 days). *Table 2* summarizes the results. The distal readings, data values received at the server in Virginia, were compared with the proximal readings, data values recorded during the phone call. For example, in the second column, systolic BP, in 109 of 126 possible instances, readings given by the participant during the phone call matched readings received automatically at AFrame servers sent wirelessly from the BP cuff. Four readings mismatched by 10% or less, and four mismatched by more than 10%. No reading was out of the expected range. In 2 cases no value was received at the AFrame servers, although a value was given over the phone, whereas five values were received at the servers but were not given over the phone (false values transmitted, or the patient may have taken a non-scheduled reading). There were 2 cases where neither a proximal nor a distal value was received although one was

**Table 2.** Comparison of Manual Readings and Values Stored in the AFrame System Shown as Number of Occurrences Within Device Column Categories

DEVICE	BP		HR		OXYGEN		WEIGHT
	SYSTOLIC	DIASTOLIC	1ST 5 SUBJECTS	2ND 4 SUBJECTS	1ST 5 SUBJECTS	2ND 4 SUBJECTS	
Both readings are the same	109	109	16	17	14	29	104
Readings within 10%	4	3	14	24	12	17	11
Readings differ by >10%	4	0	0	5	1	0	1
Reading unreasonable	0	6	0	0	8	1	0
No distal reading	2	3	29	10	22	9	8
No proximal reading	5	3	4	0	6	0	2
No distal or proximal reading	2	2	7	0	7	0	0
Total expected	126	126	70	56	70	56	126

BP, blood pressure; HR, heart rate.

expected (perhaps because of a patient not being reached on a daily phone call, or BP not being cued during a phone call).

Table 3 summarizes the reliability and usability of the third-party devices. Reliability is expressed in two ways: (1) system reliability, the number of readings in which the proximal and distal readings agree divided by the total number of proximal values; and (2) data concurrence, cases for which both proximal and distal readings are available, excluding unreasonable readings (indicating a malfunction in the device). For the purpose of the reliability analysis, readings within 10% of the manual value are considered a “correct reading” for the Nonin device because this device displays numerous values before sending a final value via Bluetooth wireless link to the CareStation application. Usability is also expressed in two ways: (1) study effectiveness, number of proximal and distal readings gathered for the device divided by the total number of expected readings; and (2) system usability, number of readings received by the system (not including unreasonable readings) divided by the number of expected readings.

On average, except for the case of the Nonin device before the firmware update, system reliability varies between 73% and 92%. Where data were available for both distal and proximal readings (data concurrence), reliability was ≥90%.

**HEALTH SURVEY DEVICE (CHUMBY)**

Data were obtained from the Chumby survey instrument on 106 of 126 days, with all eight questions answered when data were received (848/1,008). Given that only the distal values were available with this device (the daily phone call only reminded participants to fill out the questionnaire), only system usability could be assessed. Both study effectiveness and system usability compute to 84%.

**WATCH DEVICE**

*Panic button reliability.* Participants were required to push the panic button daily, but on 19 days either the participant was not

called (or was called but did not answer the phone), the experimenter failed to ask the participant to push the button during the call, or the participant failed to respond to the cue. Panic button presses were verified online on 99 out of 126 days (79% system reliability), and considering only the days where the participants were asked to push the panic button (99 out of 107 days), a 93% system usability value was obtained. To place this finding in context, consider that one study of adults older than 90 years of age who suffered falls showed that 97% failed to activate an alarm system,<sup>12</sup> suggesting that automated fall detection is probably necessary, assuming that false-positives can be minimized.

*Watch-wearing adherence.* The participant was assumed to be wearing the watch if his or her body temperature was over 76°F, based on prior calibration studies. The number of minutes wearing the watch was divided by the number of possible minutes during the study period (720 per day) to determine a value for the study use effectiveness of the watch, yielding 80%. However, the system is not currently designed to track usage or temperature when an individual leaves his or her home (it does, however, note that the watch and resident is “off campus”), so the actual system usability measurement would be higher. In fact, between the hours of 8 a.m. and 8 p.m., the watches were only in the participant’s residence (or functioning properly on battery power) 72% of the time. If the participants did not remove the watch when leaving their residences, the average daily time wearing the watch may have been as high as 10.8 h, a usability measurement of 100%. The number of minutes the watch was present in the user’s home was the number of minutes for which temperature data were available. (The watch sends temperature data to the AFrame server once per minute.)

*Overall study effectiveness and system reliability.* Study effectiveness was 88% (3,436 measurements obtained from 3,906

**Table 3. Reliability and Usability of the Third-Party Physiological Devices**

	SYSTEM RELIABILITY	DATA CONCURRENCE	STUDY EFFECTIVENESS	SYSTEM USABILITY
BP				
Systolic	0.92	0.93	0.96	0.97
Diastolic	0.90	0.97	0.96	0.91
HR				
1st 5 subjects	0.51	1.00	0.66	0.49
2nd 4 subjects	0.73	0.89	0.91	0.82
Oxygen				
1st 5 subjects	0.46	0.96	0.70	0.47
2nd 4 subjects	0.82	1.00	0.92	0.82
Weight	0.84	0.90	0.96	0.94

BP, blood pressure; HR, heart rate.

potential observations), based on nine participants  $\times$  14 days of observation  $\times$  31 variables (including proximal and distal values for systolic BP, diastolic BP, weight, pulse, and blood oxygenation level; distal values for the daily health questionnaire items 1–8; proximal values for 12 items on the comfort rating scale; and arrival of the panic alerts). Because of the large number of potential watch-worn data points, the watch study effectiveness of 80% was not included in the calculation of the overall study effectiveness.

The system reliability was calculated as an estimate of the reliability that would be expected if the readings were totally automatic, eliminating human error sources but retaining instrument and transmission error sources. For this calculation, only data sources in which both proximal and distal values were expected were included, limiting the calculation to the third-party devices and excluding the Nonin pulse oximeter for the first five participants. The system reliability value was calculated by dividing the number of data points where the distal and proximal values agreed (allowing the 10% difference for the Nonin device) by the total number of proximal readings. The combined system reliability for the third-party devices was 86%.

The overall data concurrence value was calculated to represent the percentages of correct values received by the AFrame system, by dividing the number of values at the proximal end that match values at the distal end by the total number of values not representing error codes received at the server, with distal values within 10% of the proximal value for the Nonin device considered correct. The resulting data concurrence value for the third-party devices was 94%.

Finally, the system usability was calculated to indicate how many of the expected measurements were received at the server. Readings representing error codes were not included in the number of data points received by the server because they indicate a failure either on the part of the user or with the device itself. As this calculation did not require proximal values for comparison, the Chumby data and panic button presses were included in the calculation. The watch system usability value, which may have been over 100%, was not included in the overall system usability value. As with the system reliability calculation, readings taken with the earlier Nonin device were excluded. The combined system usability value for the panic button, third-party devices, and Chumby was 87%.

## Discussion

Overall, the system was quite effective. Eighty-eight percent of the anticipated data values were gathered during the study, and 87% of the values expected to reach the AFrame server arrived. Additionally, the watch component was worn at least 80% of the requested time. Of the data intended for the AFrame server, 86% of the data from third-party devices arrived reliably at the server. Of the data that could be compared with notes taken by the study participants, 94% of the values gathered by the AFrame server agreed with the values recorded by the participants. Compared with other studies that attempted to assess different facets of reliability, our study generated comparable<sup>5</sup> or better<sup>6</sup> results, although for active components (e.g., weight, BP, oxygenation), our participants were prompted daily with

phone calls. One caveat is that payments to these healthy older participants to reimburse their time commitment may have resulted in higher motivation to adhere to study protocols than would be observed otherwise; it is also possible that unpaid but at-risk patient groups would be similarly motivated.

In summary, this study provides new, useful reliability metrics for a home-based remote monitoring system. It highlights the need to consider both passive electronic and active human components when designing such systems. On balance, the system functioned remarkably well within this population of relatively healthy older adults. In conjunction with other remote monitoring studies this study demonstrated that a well-designed system has the potential to provide significant support in the homes of older adults and others with chronic care needs. People being remotely monitored can expect to maintain residence in their own dwellings longer, potentially leading to enhanced well-being. Cost-effectiveness for deploying and training for the use of remote monitoring systems is still uncertain.<sup>13</sup> Such systems may contribute to reducing healthcare costs to the extent that they can reduce costly hospitalizations and minimize travel costs<sup>14</sup> by both patients and healthcare providers.

## Acknowledgments

This research was supported by National Institute on Aging grant SBIR 1 R43 AG029196-01A1.

## Disclosure Statement

N.C. and M.F. have no competing financial interest. A.P. and C.C. have a financial interest in AFrame Digital Inc.

## REFERENCES

- Monda J, Keipeer J, Were MC. Data integrity module for data quality assurance within an e-health system in sub-Saharan Africa. *Telemed J E Health* 2012;18:5–10.
- Lawless ST. Crying wolf: False alarms in a pediatric intensive care unit. *Crit Care Med* 1994;22:981–985.
- Tsien CL, Fackler JC. Poor prognosis for existing monitors in the intensive care unit. *Crit Care Med* 1997;25:614–619.
- Charness, N, Demiris, G, Krupinski, EA. *Designing telehealth for an aging population: A human factors perspective*. Boca Raton, FL: CRC Press, 2011.
- Osca J, Tello MJS, Navarro J, Cano O, Raso R, Castro JE, Olague J, Salvador A. Technical reliability and clinical safety of a remote monitoring system for antiarrhythmic cardiac devices. *Rev Esp Cardiol* 2009;62:886–895.
- Lee HJ, Le SH, Ha K-S, Jang HC, Chung W-Y, Kim JY, Chang Y-S, Yoo DH. Ubiquitous healthcare service using Zigbee and mobile phone for elderly patients. *Int J Med Inform* 2008;78:193–198.
- Patel S, Park H, Bonato P, Chan L, Rodgers M. A review of wearable sensors and systems with application in rehabilitation. *J Neuroeng Rehabil* 2012;9:21.
- Dhillon BS. *Reliability technology, human error, and quality in health care*. Boca Raton, FL: CRC Press, 2008.
- Kohn LT, Corrigan JM, Donaldson MS. *To err is human. Building a safer health system*. Washington, DC: National Academy Press, 2000.
- Salvendy G. *Handbook of human factors and ergonomics*, 3rd ed. Hoboken, NJ: Wiley, 2006.

11. Papadopoulos, A, Crump, C, Wilson, B. Comprehensive home monitoring system for the elderly. In: *ACM Wireless Health10 Proceedings*. San Diego, CA: ACM, 2010;214–215. doi: 10.1145/1921081.1921118.
12. Fleming J, Brayne C. Inability to get up after falling, subsequent time on floor, and summoning help: Prospective cohort study in people over 90. *BMJ* 2008;337:a2227.
13. Inglis SC, Clark RA, McAlister FA, Ball J, Lewinter C, Cullington D, Stewart S, Cleland JGF. Structured telephone support or telemonitoring programmes for patients with chronic heart failure. *Cochrane Database Syst Rev* 2010;(8):CD007228.
14. Wooton R, Bahaadinbeigy K, Hailey D. Estimating travel reduction associated with the use of telemedicine by patients and healthcare professionals: Proposal for quantitative synthesis in a systematic review. *BMC Health Serv Res* 2011;11:185.

Address correspondence to:  
Neil Charness, PhD  
Department of Psychology  
Florida State University  
1107 West Call Street  
Tallahassee, FL 32306-4301

E-mail: charness@psy.fsu.edu

Received: May 28, 2012  
Revised: September 3, 2012  
Accepted: September 20, 2012