Research Report – Research Article

A Pilot Study to Assess the Feasibility of Collecting and Transmitting Clinical Trial Data with Mobile Technologies

Colleen Russell^a Nadir Ammour^b Toby Wells^a Nicolas Bonnet^c
Matthias Kruse^a Agnes Tardat^c Christel Erales^c Thomas Shook^a
Stephane Kirkesseli^b Lionel Hovsepian^b Sy Pretorius^a

^aPAREXEL International, Waltham, MA, USA; ^bSanofi-Aventis Recherche & Développement, Chilly-Mazarin, France; ^cSanofi-Aventis Recherche & Développement, Montpellier, France

Keywords

Mobile technologies · Feasibility study · Data transmission · Agreement · Correlation · Usability · Data collection

Abstract

Background: The use of mobile technologies for data capture and transmission has the potential to streamline clinical trials, but researchers lack methods for collecting, processing, and interpreting data from these tools. **Objectives:** To assess the performance of a technical platform for collecting and transmitting data from six mobile technologies in the clinic and at home, to apply methods for comparing them to clinical standard devices, and to measure their usability, including how willing subjects were to use them on a regular basis. **Methods:** In part 1 of the study, conducted over 3 weeks in the clinic, we tested two device pairs (mobile vs. clinical standard blood pressure monitor and mobile vs. clinical standard spirometer) on 25 healthy volunteers. In part 2 of the study, conducted over 3 days both in the clinic and at home, we tested the same two device pairs as in part 1, plus four additional pairs (mobile vs. clinical standard pulse oximeter, glucose meter, weight scale, and activity monitor), on 22 healthy volunteers. **Results:** Data collection reliability was 98.1% in part 1 of the study and 95.8% in part 2 (the percentages exclude the wearable activity monitor, which collects data continuously). In part 1, 20 of 1,049 overall expected measurements were missing (1.9%), and in part 2, 45 of 1,083 were missing (4.2%). The most common reason for missing data was a single malfunctioning spirometer (13 of 20 total missed readings) in part 1, and that the subject did not take the measurement (22 of 45 total missed readings) in part 2. Also in part 2,

Nadir Ammour, DMD, MS, MBA
Sanofi-Aventis Recherche & Développement
1 avenue Pierre Brossolette
FR-91800 Chilly-Mazarin (France)
E-Mail nadir.ammour@sanofi.com

Sy Pretorius, MBChB, MMedSc, MBA, MS
PAREXEL International
195 West Street
Waltham, MA 02451 (USA)
E-Mail Sybrand.pretorius@parexel.com

a higher proportion of at-home measurements than in-clinic readings were missing (12.6 vs. 2.7%). The data from this experimental study were unable to establish repeatability or agreement for every mobile technology; only the pulse oximeter demonstrated repeatability, and only the weight scale demonstrated agreement with the clinical standard device. Most mobile technologies received high “willingness to use” ratings from the patients on the questionnaires. **Conclusions:** This study demonstrated that the wireless data transmission and processing platform was dependable. It also identified three critical areas of study for advancing the use of mobile technologies in clinical research: (1) if a mobile technology captures more than one type of endpoint (such as blood pressure and pulse), repeatability and agreement may need to be established for each endpoint to be included in a clinical trial; (2) researchers need to develop criteria for excluding invalid device readings (to be identified by algorithms in real time) for the population studied using ranges based on accumulated subject data and established norms; and (3) careful examination of a mobile technology’s performance (reliability, repeatability, and agreement with accepted reference devices) during pilot testing is essential, even for medical devices approved by regulators.

© 2018 The Author(s)
Published by S. Karger AG, Basel

Introduction

Intensifying interest in the potential of new, portable technology tools such as mobile phones and wearables to transform the conduct of clinical trials is a natural outgrowth of the broader societal trend toward digitization [1]. However, despite widespread hype and hope, few studies have tackled the substantial technical, scientific, and methodological challenges of collecting, processing, and interpreting data from mobile technologies (MTs), including wearables, in formal clinical research [2, 3].

Two powerful forces in the biopharmaceutical industry are converging to accelerate activity in this field: (1) a renewed interest in mitigating the burden on patients and investigators who participate in clinical trials of experimental agents (i.e., a focus on how to make protocols more “patient centric”) and (2) the enticing possibility that frequent (or continuous) data collection from outside the clinic will provide more study endpoints and ones that reflect somewhat less artificial study environments [4].

If mobile data gathering technologies were proven to be at least as meaningful as the current “gold standards” used in clinical trials, widespread use of these tools might, among other things, encourage greater patient participation in clinical research (due to fewer site visits), reduce the costs of conducting studies, and enable sponsors to gather more data to generate evidence about treatment performance in real-world settings [5, 6].

In this study, Sanofi and PAREXEL jointly sponsored the evaluation and technical implementation of MTs that might be suitable for extended use as additional patient evaluation tools in both pre- and post-approval clinical studies. In the context of this work, the term “mobile technology” applies to a medical device or activity tracker that has a connectivity feature allowing recorded data to be transmitted wirelessly to a central server. The study objectives were (1) to assess the reliability of a technical platform for the collection and transmission of data from multiple MTs in the clinic and at home, (2) to explore analytical methods for comparing them to clinical standard devices (CSDs), and (3) to measure their usability, including how willing subjects were to use them on a regular basis.

Fig. 1. Mobile technology and clinical standard device pairs tested in part 1 and part 2 of the study.

Subjects and Methods

Because this was a pilot feasibility study to vet the operational aspects of data collection and transmission and to implement statistical methodologies for assessing the repeatability and agreement of MTs versus CSDs, we determined that a test on closely supervised healthy volunteers should precede testing in a patient population. Part 1 of the study was a 3-week, integrated add-on to the multiple ascending dose (MAD) portion of a Sanofi phase I first-in-human study on healthy volunteers. The add-on tested two MTs, a blood pressure monitor [7] and a spirometer [8], on 25 subjects. Measurements were taken in an early phase unit (EPU) clinical setting at pre-defined times (according to the protocol) first by the CSD, then by the MT (Fig. 1; Table 1).

Part 2 was a 3-day extension study which tested six MTs (the same two as in part 1 plus four additional MTs: a pulse oximeter [9], a glucose meter [10], a weight scale [11], and a wrist-worn activity monitor [12]) on 22 subjects (Fig. 1; Table 1). A crossover design aspect was implemented to randomly assign subjects in part 2 so that half of them would have the MT data captured first followed by the CSD data, and the other half would have the CSD data captured first, followed by the MT data. Participation in part 2 was optional; a total of 18 subjects from part 1 elected to participate, while another 4 were recruited from the parent Sanofi study to participate in part 2 only. Thirteen subjects were randomly assigned to be

Table 1. Mobile technologies and clinical standard devices used in the pilot feasibility study – including the regulatory approval status

Type	Blood pressure monitor	Spirometer	Pulse oximeter	Glucose monitor	Weight scale	Activity monitor
<i>Mobile technologies</i>						
Manufacturer	A&D	Vitalograph	Nonin	Entra	A&D	Striiv
Model	UA-767PBT-Ci ¹	asma-1 bt ¹	Onyx II 9560	MyGlucoHealth	UA-351PBT-Ci	Fusion
FDA 510(k) ²	Yes	Yes	Yes	Yes	Exempt	No ⁵
CE mark ³	Yes	Yes	Yes	Yes	Yes	Yes
<i>Clinical standard devices</i>						
Manufacturer	General Electric Healthcare	Carefusion	General Electric Healthcare	Abbott ⁴	Seca	Philips Respironics
Model	Dinamap Procare 400 ¹	Masterscope ¹	Dinamap Procare 400	Freestyle Precision	930	Actiwatch
FDA 510(k) ²	Yes	Yes	Yes	Yes	Exempt	Yes
CE mark ³	Yes	Yes	Yes	Yes	Yes	Yes

¹ These mobile technologies and clinical standard devices were used both in part 1 (add-on part) and part 2 (extension part), while the other four types were used in part 2 only. ² A 510(k) is a premarket submission made to the US FDA to demonstrate that the device to be marketed is at least as safe and effective – that is, substantially equivalent – to a legally marketed device (21 CFR 807.92(a)(3)) that is not subject to premarket approval. ³ A CE mark is a legal requirement to place a device on the market in the European Union and indicates conformity with health, safety, and environmental protection standards for products sold within the European Economic Area. ⁴ The clinical standard is use of a glucose monitor – in this case, we used the Abbott Freestyle Precision glucose monitor – and laboratory results. ⁵ The Striiv Fusion is a consumer grade activity monitor, not medical grade, and thus is not subject to FDA review or clearance.

housed in the EPU for 3 days during the extension study, while 9 subjects were assigned to leave the EPU site on the afternoon of day 1 (after the last data measurement had been taken) and did the remaining measurements at home using MTs only. In the extension study, four distinct physical tests were conducted to simulate some typical clinical tests and to vary the conditions (such as rest, low activity, high activity) in which data were collected from the MTs (Table 2). All MTs carried the CE mark (a legal requirement to place a device on the market in the European Union), and all MTs except for the Striiv activity monitor and A&D weight scale had received a 510(k) Substantial Equivalence determination from the FDA (Table 1).

Data from MTs – both in the clinic and at home – were transmitted via a small plug-and-play radio communication unit, the Qualcomm Life 2net Hub (2net Hub) [13], which collected data over Bluetooth and delivered them to a Qualcomm Life 2net Cloud Platform [14] via a cellular network for storage. Platform performance was evaluated with regard to the following metrics: data collection reliability (expected vs. actual number of clinical measurements), missing data (the incidence of and reasons for missed readings), data transmission times (the time between a measurement being taken and the data arriving at the 2net Hub), and data processing times (the time between measurement data being received at the 2net Hub and the data appearing in the PAREXEL results database). The time from a measurement being taken to its appearance in the database is referred to as the end-to-end data flow time.

No reminders were sent to the subjects electronically to take readings, but while the subjects were in the EPU they were reminded to take readings by the EPU staff. For the at-home readings, no reminders of any kind were given. This was done intentionally so that the study could assess the impact of the absence of reminders.

Table 2. Summary of measurements, mobile technologies (MTs), clinical standard devices (CSDs), and physical tests conducted during part 2 of the study (extension)

Measurement	MT	CSD	Physical test
Blood pressure	A&D Model UA-767PBT-Ci ¹	GE Healthcare Dinamap Procare 400 ¹	Modified orthostatic challenge
Spirometry	Vitalograph asma-1 bt ¹	Carefusion Masterscope ¹	8-min exercise on bicycle ergometer
Oximetry	Nonin Onyx II 9560	GE Healthcare Dinamap Procare 400	Voluntary hyperventilation provocation
Glucose	Entra MyGlucoHealth	Abbott Freestyle Precision (plus lab result)	Meal test
Weight	A&D Model UA-351PBT-Ci	Seca 930	Not applicable
Activity	Striiv Fusion	Philips Respironics Actiwatch	8-min exercise on bicycle ergometer

GE, General Electric. ¹ These MTs and CSDs were used both in part 1 (add-on part) and part 2 (extension part), while the other four were used in part 2 only.

This was not a formal MT validation study. However, an important secondary objective was to gain experience in comparing measurements taken by MTs and those taken by CSDs. For five of the MT-CSD pairs, agreement was assessed using two statistical methods: (1) the Bland-Altman approach [15] and (2) a modification of this approach by Francq and Govaerts [16]. Clinical acceptance ranges (CARs) were defined prior to data analysis (Table 3) [17–20]. The activity monitors were assessed for correlation only (vs. agreement), because the MT activity watch and the CSD activity watch measured different endpoints [21].

The MT population included all data transmitted, whereas the per protocol population excluded invalid device readings. In this study, invalid device readings were defined as data points that were (1) erroneous in some way due to the use of the device itself (e.g., a family member steps on the weight scale at home and it transmits data from someone who is not a study subject) or (2) physically impossible (such as an increase in FEV₁ lung capacity of 2 L/s) and therefore not valid due to device malfunction. A small number of spare devices were available for use in case of malfunction. If a subject noticed an issue with a device, they could raise a concern with the EPU staff. They in turn would alert technical staff and a substitution was made. If a subject did not notice an issue, no intervention was made. All MTs had “store-and-forward” capability, whereby a reading which could not immediately be sent would be stored on the device until the next available opportunity.

Because the research was exploratory, no statistical hypothesis was specified, no power calculations were conducted to determine the sample size, and no *p* values are reported. Various scenarios were explored for some devices and study parts (e.g., analyses based on data from part 1 days 1–14 pre-dose, or analyses based on data from part 2 tests). The devices were assessed for repeatability (the amount of variation within one subject on the same device under the same conditions) and agreement (the amount of variability and accuracy between the device pair relative to the CAR). Due to the exploratory nature of the study, defin-

Table 3. Clinical acceptance ranges

Clinical endpoint	Device comparison (MT vs. CSD)	Range (lower, upper)
Blood pressure, mm Hg	A&D Model UA-767PBT-Ci vs. GE Healthcare Dinamap Procare 400	Level 1: (–5, 5) Level 2: (–10, 10) Level 3: (–15, 15)
Pulse, bpm	A&D Model UA-767PBT-Ci vs. GE Healthcare Dinamap Procare 400	Level 1: (–5, 5) Level 2: (–10, 10)
FEV ₁ , L	Vitalograph asma-1 bt vs. Carefusion Masterscope	Level 1: (–5, 5) Level 2: (–10, 10)
Glucose, mmol/L	(1) Entra MyGlucoHealth vs. Abbott Freestyle Precision (2) Entra MyGlucoHealth vs. laboratory result (3) Abbott Freestyle Precision vs. laboratory result	(–0.67, 0.67)
Blood oxygen, %	Nonin Onyx II 9560 vs. GE Healthcare Dinamap Procare 400	(–3, 3)
Weight, kg	A&D Model UA-351PBT-Ci vs. Seca 930	(–1, 1)
MT, mobile technology; CSD, clinical standard device; GE, General Electric.		

itive conclusions about device interchangeability were not made. Nevertheless, the data support general observations on device performance and recommendations for future study designs.

The subjects and investigators using the MTs evaluated the usability (“ease of use”) with a Subject Device Usability Questionnaire (DUQ) [22] and Investigator After-Scenario Questionnaire (ASQ) [23], respectively.

Results

Healthy volunteers were enrolled at a PAREXEL EPU between February 11 and June 13, 2017. In part 1 of the study, data collection reliability across the two devices was high: 98.1% of all expected readings were taken at the correct time and processed successfully. Compliance was slightly higher for the blood pressure monitor at 99.2% versus 96.5% for the spirometer. In part 2 of the study, data collection reliability across all devices (excluding the activity monitor) was also high at 95.8%. The devices with the highest compliance were the blood pressure monitor (98.9%) and spirometer (98.1%), followed by the weight scale (93.8%), pulse oximeter (89.9%), and glucose meter (88.9%). The activity monitor collected data continuously, so compliance was calculated differently. Of a total of 66 expected data files, 55 (86%) were received and read successfully.

In part 1 of the study, 20 of 1,049 expected readings were missing. The majority of these (13) were due to a single spirometer malfunctioning. Five others were due to subjects not taking the reading, or not taking it at the correct time. Equipment malfunction caused the remaining 2 missing readings. None were due to platform processing issues. In part 2 of the study, of 1,083 expected measurements, 45 were missing (4.2%). By far the most common reason for a missing measurement was that it was not taken (22; 49%), which occurred for

Table 4. Mobile technologies listed by repeatability, agreement, and user acceptance

Device type and model	Repeatability ¹	Agreement ¹	User acceptance ²
Weight scale			
A&D model UC-351PBT-Ci	Not within CAR	Within CAR	21/22 (95%)
Pulse oximeter			
Nonin Onyx II 9560	Within CAR	Not within CAR	16/16 (100%)
Spirometer			
Vitalograph asma-1 bt	Not within CAR	Not within CAR	19/22 (86%)
Blood pressure monitor			
A&D model UA-767PBT-Ci	Not within CAR	Not within CAR	18/22 (82%)
Activity monitor			
Striiv Fusion	n/a	n/a	12/22 (55%)
Glucose meter			
Entra MyGlucoHealth	Not within CAR	Not within CAR	16/22 (73%)

CAR, clinical acceptance range. ¹ Repeatability and agreement were assessed in relation to the CARs shown in Table 3. ² Willingness to use the device more than once per day for more than 1 month (percentage of subjects answering “agree” or “strongly agree”). See online supplementary material for more information regarding user acceptance data.

all device types. User error with the glucose meter accounted for the second largest category (15; 33% of all missing measurements), and error messages with the pulse oximeter accounted for the third largest (4; 9%). For the activity monitor there were two causes of missing data: the battery on a single device was drained and not recharged (5%), and some devices were turned off before final data transmission on day 3 (9%). All causes of missed measurements were at the device and data collection level; no data were lost in the data processing stages. At-home measurements accounted for a higher proportion of missed measurements than in-clinic readings in part 2 of the study (12.6 vs. 2.7%).

In both parts of the study, the transmission times (the time between a measurement being taken and the data arriving at the 2net Hub) for each device could be split into two categories: fast (when data flow encountered no interruptions) and slow (when there were connectivity issues), the boundary between the two being device dependent. The store-and-forward capability of the MTs studied proved very important, because the volume of messages with “slow” transmission times due to connectivity issues was higher than expected. There was no significant evidence that the transmission times of the measurements taken at home were different from those taken in the clinic. The processing times (the time between measurement data being received by the 2net Hub and the data appearing in the PAREXEL results database) were similar across the device types, whether at home or in the clinic.

An analysis of the end-to-end data flow times showed that in part 2 of the study, the measurement data arrived in the results database a median of 34 s after being taken. Overall, 90% of the measurements appeared in the results database within 11 min 9 s of being taken.

General observations on repeatability, agreement, and user acceptance for all six MTs are summarized in Table 4. This study was not able to establish repeatability or agreement for every MT (see online suppl. Material; see www.karger.com/doi/10.1159/000493883 for all online suppl. material). One MT (the pulse oximeter) demonstrated repeatability within the CAR and one device (the weight scale) demonstrated agreement (vs. the CSD) within the CAR. Below is a discussion of each MT.

Fig. 2. Plot of systolic blood pressure comparing mobile technology (MT) to clinical standard device (CSD) pre-dose on days 1–14 with the $x = y$ line of identity during part 1 of the study. Each symbol represents an individual's measurement pairs over all time points.

Fig. 3. Plot of systolic blood pressure (SBP) difference versus average SBP with limits of agreement (LoA) and tolerance intervals (TI) pre-dose on days 1–14 during part 1 of the study. “x” = paired differences at planned time points; “o” = by-subject average difference. MT, mobile technology; CSD, clinical standard device.

A&D Wireless Blood Pressure Monitor UA-767PBT-Ci

In the part 1 scenario based on one pre-dose measurement from each device on days 1–14, the systolic blood pressure (SBP) repeatability coefficient for the MT was 22.4 mm Hg, that is, on average, the absolute difference between two readings for the same subject using the MT under similar conditions is expected to be less than 22.4 mm Hg for 95% of the subjects. The SBP results with the CSD were similar (repeatability coefficient = 22.6 mm Hg, excluding an outlier), and both exceeded the highest level of CAR of 15 mm Hg, a reflection of the amount of variability in one SBP measurement.

The difference between the SBP devices in pre-dose assessments from days 1–14 collected at the same time point can be observed with respect to the line of identity ($x = y$) in Figure 2. Figure 3 shows that both the limits of agreement (LoA) and tolerance intervals are similar to each other but outside the level 3 CAR (Table 3). The LoA/tolerance interval describes the interval within which the difference between an MT and a CSD (based on one measurement) can be expected to lie 95% of the time. The intervals are calculated from the individual standard deviation (SD) estimates, which are a combination of the between- and within-subject variances. In this scenario, the MT appeared to be accurate with respect to the CSD (mean difference, -0.3 mm Hg), but the variability for both devices led to a failure to establish agreement based on one measurement per time point. In a post hoc analysis considering the hypothetical situ-

Fig. 4. Systolic blood pressure (SBP) correlated bivariate least squares regression for the average of three observations post hoc during part 1 of the study. “○” = by-subject average difference. MT, mobile technology; CSD, clinical standard device; PI, prediction interval based on 1 future observation; GI, prediction interval based on the average of 3 future observations; CB, confidence band for the regression line.

ation based on the average of three measurements at each time point, the predictive intervals from correlated bivariate least-squares regression fell within level 3 CAR when SBP readings were around 110 mm Hg (Fig. 4). The diastolic blood pressure findings from part 2 morning readings for in-clinic subjects compared to level 1 CAR were consistent with SBP, although a small bias was observed (2.5 mm Hg). These data confirm that one blood pressure reading is highly variable, may not be representative of the true underlying blood pressure, and may make it difficult to establish agreement between devices. In addition, since blood pressure and pulse have different magnitudes and relative variability, the behavior of each biometric from the MT relative to the CSD should be fully understood if planned for use in a clinical trial.

Vitalograph asma-1 bt Spirometer

The mean difference in both parts of the study was a negative bias, with more MT results being lower than the respective CSD results. FEV₁ is a subject-dependent, effort-driven clinical parameter. Since the CSD in part 1 was used first at each measurement time and multiple FEV₁ samples were collected in the MAD study (up to 8 trials), it was hypothesized that subject fatigue could have contributed to the lower MT readings. In part 2, the variability and bias were instead attributed to the clinical setting and conduct of sample collection. For the CSD, site staff accompanied the subjects and encouraged them to blow hard into the device in order to capture the maximum of three trials at each reading. For the MT, the subjects were provided some instruction and then left alone to conduct the three trials to simulate a home environment. These factors likely contributed to the bias and lack of repeatability.

Nonin Onyx II 9560 Pulse Oximeter

The Nonin Onyx II 9560 demonstrated repeatability in the scenario where three morning readings were taken by the subjects who took the device home. For the hyperventilation test, the LoA were outside of but close to the CAR. The narrow range of results from the healthy population and the upper limit of 100% may have been factors; an alternative data transformation, or testing the device in the disease population, could clarify this issue.

Entra MyGlucoHealth Glucose Meter

The Entra MyGlucoHealth measurements were consistently higher than both the standard glucose meter and the laboratory results (Fig. 5). Although these data were collected during a meal test (collected before a meal in a fasted state and after the meal at 30, 60, 90, and 120 min), the high degree of bias and variability was unexpected.

Fig. 5. Plot of glucose comparing mobile technology (MT) to laboratory results (CSL) from the meal test with the $x = y$ line of identity during part 2 of the study (extension).

A&D Weight Scale UA-351PBT-Ci

Due to the nature of day-to-day weight fluctuations, repeatability over three morning measurements was high at 1.5 kg for both devices; i.e., two measurements taken from the same subject on the same device are expected to be within 1.5 kg for 95% of the subjects on average. To assess agreement, day-to-day variability in weight was accounted for by using the difference between daily morning measurements from both devices. The LoA were within the CAR of ± 1.0 kg, and the difference between the two devices was negligible (-0.09 kg).

Striiv Fusion Activity Watch

This version of the Striiv Fusion is primarily a step counter; it captures actual steps above a level of intensity (in contrast, the CSD, Actiwatch, can measure small motor activity). Motion such as cycling cannot be accurately counted as a step and will result in false-positive data. Thus, the study was not able to measure the correlation between total steps per hour from the Striiv and total activity per hour from the Actiwatch.

Usability Questionnaires

Most MTs registered high “willingness to use” ratings (defined as an answer of “agree” or “strongly agree” from subjects when asked whether they would be willing to use the device more than once per day for more than 1 month) in the DUQ (Table 4; see online suppl. material). The ASQ scores indicate that the investigators (5 total in part 1 and 2 in part 2) expressed moderate-to-high satisfaction with the blood pressure monitor and spirometer and did not express dissatisfaction with any of the devices. The DUQ ratings collected on day 3 indicate that the majority of subjects expressed moderate-to-high “overall” satisfaction with all of the devices.

Discussion/Conclusion

MTs should not be used in clinical research unless the wireless data flow and processing platform are dependable. This study demonstrated that the technical platform produced reliable and timely results, with little data loss (the minimal losses were primarily due to measurements not taken, user error, and malfunctioning devices). The device instructions for subjects and site training contributed to a high level of data collection reliability.

That 90% of the data were delivered to the main PAREXEL results database within 12 min demonstrates the potential of marrying MTs with wireless transmission to a clinical trial database. Nevertheless, the study highlighted potential problems with at-home data collection. Although most readings were actionable within a minute, connectivity issues between the MTs and transmission devices caused significant delays to this. There was also significant variability in transmission time between devices. Both these factors must be considered when developing real-time algorithms. No active data monitoring occurred during the study, which was a factor contributing to a large proportion of the missed readings. Feedback to the subject when expected readings had been missed may have prevented future missed readings due to forgetfulness or triggered an investigation that would identify a fault more quickly. Equally, monitoring error messages from a device or determining when it has a low battery level could lead to interventions that prevent missed readings. Thus, this study has highlighted the importance of a comprehensive device and data monitoring system which feeds back to the subject.

There are limitations to the generalizability of our data. We studied healthy volunteers and results may not be the same with actual patients. Our sample sizes may not have been adequate to account for variables such as age, ethnicity, socioeconomic status, or educational level. Finally, since the study participants were compensated, they may have been less likely to express dissatisfaction with a device in the questionnaire.

The process of comparing the data from the MTs with those from the CSDs identified at least three critical areas of study for improving methods that will enable MTs to streamline clinical trials.

First, if an MT captures more than one type of endpoint (such as blood pressure and pulse), agreement between the MT-CSD device pair may need to be established for each endpoint to be included in a clinical trial. Specifically for the use of blood pressure monitors, we recommend more than one blood pressure reading per time point, because blood pressure is highly variable and one reading may not be representative of the true underlying blood pressure.

Second, researchers need to develop criteria for excluding invalid device readings from the intent-to-treat population using ranges based on accumulated subject data and established norms. To handle data that may not be valid, algorithms to identify suspect readings will need to be set up prospectively, preferably via the data transmission platform. It will be essential to have mobile phone apps that identify potentially erroneous and missing data and send prompts to subjects or patients in real time. Alternatively, a companion app could offer to the patient the ability to mark erroneous data by flagging them.

Finally, a key finding of this study is that documented performance of MTs (i.e., reliability, repeatability, and agreement with accepted reference devices) during pilot testing turned out to be essential, even for 510(k)-approved MTs and those with a CE mark, to set expectations for use in clinical trials [6]. Study designs could then account for all known factors. MTs which fail to demonstrate reliability, repeatability, and agreement with reference devices for appropriately selected endpoints may need to be recalibrated or replaced with alternate MTs.

Acknowledgements

Jean-Marc LeBideau provided the technical coordination and integration of platform technology with test devices. Ramona Borst served as Clinical Research Coordinator for Sanofi MAD study conduct and device measurements. Kay Reichenbach served as Clinical Research Coordinator for pilot study conduct, device measurements, and contribution to the development of the platform. Laurent Venerucci and Jens Janiszewski helped with device

selection and testing, preparation of device kits, and allocation to subjects. Clifford Mbonbowo integrated data across platforms and created data sets. Alain Afios served as the Information Technology Solution Architect, helping with the review and design of the data flow, data collection, and data transmission.

Statement of Ethics

The subjects (or their parents or guardians) enrolled in this study have given their written informed consent. Prior to the initiation of the study, the final clinical study protocol, subject information sheet, and informed consent form were approved by the Ärztekammer Berlin Ethics Committee.

Disclosure Statement

C.R., T.W., M.K., T.S., and S.P. are employed by PAREXEL International. N.A., N.B., A.T., C.E., S.K., and L.H. are employed by Sanofi-Aventis Recherche & Développement. The authors have no other conflicts of interest to declare.

Funding Sources

Both Sanofi-Aventis Recherche & Développement and PAREXEL funded the study and covered costs related to protocol development, study execution, and data collection and analysis. More than 50% of the direct and indirect costs were absorbed by PAREXEL.

Author Contributions

C.R. contributed to the secondary endpoint data analysis and reviewed the manuscript. T.W. contributed to the primary endpoint data analysis and reviewed the manuscript. T.S. contributed to the study design and reviewed the manuscript. S.P. reviewed the manuscript. M.K. was involved in the setup and design of the study, analyzed data, contributed to the original clinical study report, and reviewed the manuscript. N.A. contributed to the launch of this project and development of the study protocol and reviewed the final report. N.B. contributed to the development of the study protocol and statistical analysis plan and reviewed the study report. C.E. helped coordinate the conduct of the study and the collaboration with PAREXEL and reviewed the study report. A.T. acted as clinical trial manager of the phase I study that provided the subjects and clinical standard data for comparison to the wearable device data. S.K. contributed to the development of the study protocol and reviewed the final report. L.H. contributed to the review of the study report.

References

- Munos B, Baker PC, Bot BM, Crouthamel M, de Vries G, Ferguson I, et al. Mobile health: the power of wearables, sensors, and apps to transform clinical trials. *Ann N Y Acad Sci*. 2016 Jul;1375(1):3–18.
- Izmailova ES, Wagner JA, Perakslis ED: Wearable Devices in Clinical Trials: Hype and Hypothesis. *Clin Pharm Ther*. 2018 Jul;104(1):42–52.
- Perry B, Herrington W, Goldsack JC, Grandinetti CA, Vasisht KP, Landray MJ, et al. Use of Mobile Devices to Measure Outcomes in Clinical Research, 2010–2016: A Systematic Literature Review. *Digit Biomark*. 2018 Jan;2(1):11–30.
- Banegas JR, Ruilope LM, de la Sierra A, Vinyoles E, Gorostidi M, de la Cruz JJ, et al. Relationship between Clinic and Ambulatory Blood-Pressure Measurements and Mortality. *N Engl J Med*. 2018 Apr;378(16):1509–20.
- Byrom B, Watson C, Doll H, Coons SJ, Eremenco S, Ballinger R, et al.; ePRO Consortium. Selection of and Evidentiary Considerations for Wearable Devices and Their Measurements for Use in Regulatory Decision Making: Recommendations from the ePRO Consortium. *Value Health*. 2018 Jun;21(6):631–9.
- Clinical Trials Transformation Initiative [Internet]. CTTI Recommendations: Advancing the Use of Mobile Technologies for Data Capture and Improved Clinical Trials. July 16, 2018 [accessed Sep 9, 2018]. Available from: <https://www.ctti-clinicaltrials.org/sites/www.ctti-clinicaltrials.org/files/mobile-devices-recommendations.pdf>.
- A&D Medical [Internet]. Premium Wireless Blood Pressure Monitor Model UA-767PBT-Ci [accessed March 22, 2018]. Available from: <https://medical.andonline.com/product/premium-wireless-blood-pressure-monitor/ua-767pbt-ci>.
- Vitalograph® [Internet]. asma-1 btTM [accessed March 22, 2018]. Available from: <https://vitalograph.com/product/162431/asma-1-bt>.
- Nonin Medical [Internet]. Onyx® II 9560 [accessed March 22, 2018]. Available from: <http://www.nonin.com/Onyx9560>.
- Entra Health Systems [Internet]. MyGlucoHealth® Wireless Meter [accessed March 22, 2018]. Available from: <http://www.myglucohealth.org/wireless.html>.
- A&D Medical [Internet]. UA-351PBT-Ci Wireless Weight Scale with Bluetooth Communication [accessed March 22, 2018]. Available from: <http://www.andmedical.com.au/products-service/telemedicine-products/uc-321pbt-scale-with-bluetooth-communication-2>.
- Striiv [Internet]. Fusion Activity Tracker + Smartwatch [accessed March 22, 2018]. Available from: <https://www.striiv.com/pages/striiv-fusion>.
- Qualcomm Life [Internet]. 2netTM Hub [accessed March 22, 2018]. Available from: <https://qualcomm-life.com/2net-hub>.
- Qualcomm Life [Internet]. 2netTM Platform [accessed March 22, 2018]. Available from: <https://qualcomm-life.com/2net-platform>.
- Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999 Jun;8(2):135–60.
- Francq BG, Govaerts B. How to regress and predict in a Bland-Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. *Stat Med*. 2016 Jun;35(14):2328–58.
- O'Brien E, Pickering T, Asmar R, Myers M, Parati G, Staessen J, et al.; Working Group on Blood Pressure Monitoring of the European Society of Hypertension. Working Group on Blood Pressure Monitoring of the European Society of Hypertension International Protocol for validation of blood pressure measuring devices in adults. *Blood Press Monit*. 2002 Feb;7(1):3–17.
- Barr RG, Stemple KJ, Mesia-Vela S, Basner RC, Derk SJ, Henneberger PK, et al. Reproducibility and validity of a handheld spirometer. *Respir Care*. 2008 Apr;53(4):433–41.
- Lacara T, Domagtoy C, Lickliter D, Quattrocchi K, Snipes L, Kuszaj J, et al. Comparison of point-of-care and laboratory glucose analysis in critically ill patients. *Am J Crit Care*. 2007 Jul;16(4):336–46.
- U.S. Food and Drug Administration [Internet]. Pulse Oximeters – Premarket Notification Submissions [510(k)s]: Guidance for Industry and Food and Drug Administration Staff. March 4, 2013 [accessed March 22, 2018]. Available from: <https://www.fda.gov/RegulatoryInformation/Guidances/ucm341718.htm>.
- Hamlett A, Ryan L, Serrano-Trespacios P, Wolfinger R. Mixed models for assessing correlation in the presence of replication. *J Air Waste Manag Assoc*. 2003 Apr;53(4):442–50.
- Lewis JR [Internet]. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. Technical Report 54.786. 1993 [accessed March 22, 2018]. Available from: <http://cite-seerx.ist.psu.edu/viewdoc/download?doi=10.1.1.584.6610&rep=rep1&type=pdf>.
- Lewis JR. Psychometric evaluation of the PSSUQ using data from five years of usability studies. *Int J Hum Comput Interact*. 2002;14(3):463–88.