Published in final edited form as:
Am J Prev Med. 2013 August ; 45(2): 228–236. doi:10.1016/j.amepre.2013.03.017.

Mobile Health Technology Evaluation:

The mHealth Evidence Workshop

Santosh Kumar, PhD, **Wendy J. Nilsen, PhD**, **Amy Abernethy, MD**, **Audie Atienza, PhD**, **Kevin Patrick, MD, MS**, **Misha Pavel, PhD**, **William T. Riley, PhD**, **Albert Shar, PhD**, **Bonnie Spring, PhD**, **Donna Spruijt-Metz, PhD**, **Donald Hedeker, PhD**, **Vasant Honavar, PhD**, **Richard Kravitz, MD**, **R. Craig Lefebvre, PhD**, **David C. Mohr, PhD**, **Susan A. Murphy, PhD**, **Charlene Quinn, PhD**, **Vladimir Shusterman, MD, PhD**, and **Dallas Swendeman, PhD, MPH**
Department of Computer Science (Kumar), University of Memphis, Memphis, Tennessee; the Office of Behavioral and Social Sciences Research (Nilsen), the National Cancer Institute (Atienza), the National Heart, Lung and Blood Institute (Riley), NIH, Bethesda, the Department of Epidemiology and Public Health (Quinn), University of Maryland School of Medicine, Baltimore, Maryland; the Department of Medicine (Abernethy), Duke University Medical Center, Durham, the Health Communication and Marketing (Lefebvre), RTI International, Research Triangle Park, North Carolina; the Department of Preventive Medicine (Patrick), University of California, San Diego, La Jolla, the Department of Preventive Medicine (Spruijt-Metz), University of Southern California, the Department of Psychiatry and Biobehavioral Sciences (Swedenman), David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, the Department of Internal Medicine (Kravitz), University of California, Davis, Sacramento, California; the Directorate for Computer and Information Science and Engineering (Pavel, Honaver), National Science Foundation, Arlington, Virginia; the Pioneer Portfolio (Shar), Robert Wood Johnson Foundation, Princeton, New Jersey; the Department of Preventive Medicine (Spring, Mohr), Northwestern University, Evanston, the Division of Epidemiology and Biostatistics (Hedeker), University of Illinois at Chicago, Chicago, Illinois; the Department of Statistics (Murphy), University of Michigan, Ann Arbor, Michigan; the Noninvasive Cardiac Electrophysiology Laboratories (Shusterman), University of Pittsburgh Medical Center & PinMed, Pittsburgh, Pennsylvania

Abstract

Creative use of new mobile and wearable health information and sensing technologies (mHealth) has the potential to reduce the cost of health care and improve well-being in numerous ways. These applications are being developed in a variety of domains, but rigorous research is needed to examine the potential, as well as the challenges, of utilizing mobile technologies to improve health outcomes. Currently, evidence is sparse for the efficacy of mHealth. Although these technologies may be appealing and seemingly innocuous, research is needed to assess when, where, and for whom mHealth devices, apps, and systems are efficacious.

In order to outline an approach to evidence generation in the field of mHealth that would ensure research is conducted on a rigorous empirical and theoretic foundation, on August 16, 2011, researchers gathered for the mHealth Evidence Workshop at NIH. The current paper presents the results of the workshop. Although the discussions at the meeting were cross-cutting, the areas covered can be categorized broadly into three areas: (1) evaluating assessments; (2) evaluating

Address correspondence to: Wendy J. Nilsen, PhD, Office of Behavioral and Social Sciences Research, NIH, 31 Center Dr., Bldg. 31, B1C19, Bethesda MD 20892. nilsenwj@od.nih.gov.

interventions; and, (3) reshaping evidence generation using mHealth. This paper brings these concepts together to describe current evaluation standards, future possibilities and set a grand goal for the emerging field of mHealth research.

Introduction

Creative use of new mobile health information and sensing technologies (mHealth) has the potential to reduce the cost of health care and improve health research and outcomes. These technologies can support continuous health monitoring at both the individual and population level, encourage healthy behaviors to prevent or reduce health problems, support patient chronic disease self-management, enhance provider knowledge, reduce the number of healthcare visits, and provide personalized, localized, and on-demand interventions in ways previously unimaginable.[1–3] In this paper, *mobile technology* is defined as wireless devices and sensors (including mobile phones) that are intended to be worn, carried, or accessed by the person during normal daily activities. As highlighted in Figure 1, mHealth is the application of these technologies either by consumers or providers, for monitoring health status or improving health outcomes, including wireless diagnostic and clinical decision support.

mHealth applications are being developed and evaluated in a variety of domains, including diabetes,[4] asthma,[5] obesity,[6,7] smoking cessation,[8] stress management,[9] and depression treatment.[10] However, whether mHealth leads to better overall health outcomes and reduced disease burden is still unknown. For example, recent studies note that short messaging service (SMS)-based health interventions have not been adequately tested for efficacy,[11] and few smoking cessation smartphone apps are evidenced based.[12] Rigorous research is needed that examines the potential, as well as the challenges, of using mobile technologies to improve health outcomes. mHealth devices, apps, and systems may be ineffective or, at worst, yield adverse outcomes on the quality or cost outcomes of health. In a healthcare system already burdened with suboptimal outcomes and excessive costs, premature adoption of untested mHealth technologies may detract from, rather than contribute to, what is needed for true overall health improvement.

In order to outline an approach to evidence generation to ensure mHealth research has a rigorous empirical and theoretic foundation, on August 16, 2011, researchers from the domestic and international community, policymakers, health professionals, technologists, and representatives from regulatory and funding agencies gathered for the invited mHealth Evidence Workshop at NIH. The meeting was sponsored by the Pioneer Portfolio of the Robert Wood Johnson Foundation, the McKesson Foundation, the National Science Foundation, and the Office of Behavioral and Social Sciences Research and the National Heart, Lung and Blood Institute at NIH. Table 1 provides a list of participants at the meeting who also contributed to the current paper, in addition to the authors listed above.

This paper presents the results of the workshop's participants' discussion summarized into opportunities and challenges in three areas of mHealth evidence generation where unique issues are emerging: (1) evaluating assessments; (2) evaluating interventions; and (3) reshaping evidence generation using mHealth. Some are issues traditionally addressed in medical or health behavior research, but others are less common in health-related research and reflect the need to borrow research methods from other domains such as engineering and the systems sciences. The final section of the paper addresses other key issues for mHealth research.

Evaluating Assessment

mHealth technologies support new methods for collecting biological, behavioral, or environmental data and the outcomes of interventions. These include sensors that monitor phenomena with higher precision, improved sampling frequency, fewer missing data, greater convenience, and in some cases, lower cost than traditional measures. Algorithms derived from sensor data and self-reports allow inferences about physiologic, psychological, emotional, and environmental state, such as mobile sensor systems for psychological stress[9] or smoking.[8]

Reliability and Validity

As with any measure, before mHealth assessment methods can be recommended, their reliability and validity must be established.[13] Table 2 highlights the goals and challenges of attaining reliability and validity in mHealth assessments, not the least of which is the rapid evolution of device technologies that may affect the quality of data they produce and thus influence their reliability and validity.

The very nature of reliability assessment in free-living samples—across subjects, times of day, or days of the week—challenges the value of reliability estimates derived from studies conducted in controlled laboratory environments. Research is needed to better understand the effect of the variability on collecting time-intensive data in real-world settings. mHealth devices are frequently used by individuals with little training, or in situations where comfort and convenience are paramount. For example, wearable device placement may need to be negotiable even though it may affect data quality. Novel approaches to determining reliability are needed that incorporate factors such as the impact of placement changes and data collection models that do not have the same pristine information flows that could be collected in the lab (e.g., using a mobile phone microphone to assess sound).

Establishing validity usually requires existence of gold standards that measure the same or similar constructs. mHealth now enables us to use common measures across the population and around the globe, which means that a single construct may not have one gold standard. For example, walking may not have the same data signature in an elderly person as in a child. Thus, the gold standard for walking will have to be based on formulas that take into account variability within the population and environment.

Intervention Evaluation

Matching the Rapid Pace of mHealth with Existing Research Designs

Evidence requirements for new interventions in health are well established. Experiments are conducted to evaluate the efficacy and effectiveness of new treatments and prevention programs. The RCT has long been the gold standard for research for determining the efficacy of health interventions.[14] However, RCTs have a long time lag (i.e., 5.5 years on average) from the initiation of subject recruitment to publication of the outcome.[15] In addition, RCTs pose additional challenges due to cost, randomization for treatment assignment, and/or the level of treatment adherence required.[16]

In mHealth, this time lag is critical because the technology may be obsolete before the trial is completed. In some cases, the rapidly evolving nature of both mHealth technologies and their uptake in the population and healthcare settings mean that some components of mHealth interventions may need to continuously improve during a trial. This gap may lead developers to move quickly from pilot to dissemination or skip outcome evaluations altogether to avoid a full-scale RCT, threatening understanding of the long-term value of mHealth.

Recent work in mHealth, and the data "revolution" that it augers, suggests that mHealth's capabilities may change the strengths, weaknesses, and feasibility of existing research methods and may even enable development of new, more efficient designs. Many open questions exist about use of current methods in mHealth research. For example, can obtaining multiple repeated measures on a few participants, rather than a few measures on many participants, reduce the size of clinical trials and render conventional research designs, such as RCTs, more efficient (i.e, quicker, cheaper, and more viable for the rapidly moving technology-based mHealth in terms of time and recruitment)? In this section, several potential research designs to evaluate the efficacy and effectiveness of mHealth interventions are described. Table 3 outlines these designs, grouped by stage of development of the mHealth intervention.

Accommodating Continuously Evolving Technology in Research Designs

To address the problem noted above about mHealth technologies becoming obsolete before they are fully tested, researchers might wish to provide upgrades on a regular basis. However, this runs counter to scientific norms where changes to an intervention during a research study threaten internal validity. To address this issue, continuous evaluation of evolving interventions (CEEI) has been proposed as one method for testing evolving mHealth interventions.[22] In CEEI, substantively new versions are deployed along with the previous version, with users randomized to available versions. The most efficacious version, based on a priori criteria, is retained. The CEEI design may also be well suited to ongoing evaluation of interventions as they go to scale, continuously improve over time, and adapt to rapidly changing technologies.

Although the CEEI allows a fine-grained level of testing and inference around specific design features, the traditional RCT may still be applied to mHealth interventions if the level of inference is made around a package of robust intervention features or functions whose delivery will naturally adapt to changing technology environments and preferences over time and across contexts during dissemination. The integration of assessment and intervention methods in mHealth also holds the potential to make ongoing and continuous evaluation feasible and cost effective, as well as to improve design.

Model-Based Design of Adaptive Interventions Using mHealth

One of the promises for mobile technologies is the potential to use them to tailor and personalize interventions in real time. This may lead to adaptive interventions that reduce waste, increase compliance, and enhance the potency of an intervention.[23,24] To accomplish this, however, will require a better understanding of within-subject differences and the effects of hypothesized mediating variables on outcomes. Thus, statistical methods that better specify within- and between-subjects effects are needed. For example, knowing if variations in mood are related to health behaviors such as smoking, eating, or exercise may be critical for tailoring and personalizing interventions.

Recent developments have begun addressing tailoring and intervention optimization in treatment research. In the multiphase optimization strategy (MOST), promising components of an intervention are identified in a screening phase through either factorial or fractional factorial analysis of variance design. These promising components can then be evaluated in a confirmatory randomized trial. For refining the intervention, sequential multiple assignment randomized trial (SMART) can be used where individuals are randomly assigned to various intervention choices over time. In SMART, researchers decide which aspects of treatments require investigation and then randomize individuals at each treatment decision based on feasibility, ethical issues, or other factors.[24]

Reshaping Evidence Generation Using mHealth

mHealth technologies also offer new capabilities for evaluating the efficacy of both traditional and mHealth interventions while reducing the time and resources needed. Several of these (described below), when combined with the statistical enhancements, such as modeling and machine learning, will enable improvements in the speed and efficiency of evaluation.[26] These advantages reflect fundamental scientific issues that set mHealth apart from the traditional approaches.

High Data Density

Mobile technologies can provide data at very high sampling rates (e.g., 10–500 times per second) that support the quantification of phenomena (e.g., physical activity) that previously was only poorly understood because of intermittent and limited measurement. The high density of data in conjunction with time-series analysis can increase the discriminative power of any experimental design. High-density data can also facilitate exploration of subtle patterns or “fingerprints” that may better explain intervention or treatment effects in shorter intervals than previous methods.[27] Further, such intensive longitudinal data can allow one to examine effects on variances, both between- and within-subjects, as well as on mean levels of parameters of interest.[28]

Data Processing and Analytics

High data density requires data processing methods not commonly used in health research. Machine-learning methods that make classification decisions based on features from the data can be applied to segments of data to draw inferences about individuals such as type of physical activity, level of stress, or intensity of pain. Having accurate analytics for high-frequency data collected in mHealth applications is critical for both assessment and intervention purposes.

Examples of methods for classification decisions include: unsupervised cluster analysis, latent class analysis, and latent semantic analysis; more-complex models such as topic models are used to learn associations and patterns using numeric or textual data.[29] If training data are available, supervised classification algorithms are more efficient. Another popular classification technique, support vector machines (SVM) can lead to robust classification performance, but in some cases, simpler algorithms such as decision trees may also provide sufficient classification accuracies.[30]

In many situations, the variables of interest cannot be observed directly and must be inferred from those that are directly measurable. In those cases the inferences must be made using various model-based techniques, for example hidden Markov models.[31] In a similar fashion, factor analysis and latent trait models can be used to reveal the internal structure of the data in a way that best explains the variance among the measured variables or items.[32]

What these processes share is a goal to minimize bias and achieve high predictive accuracy of a data classifier. This is a multi-step process that: (1) defines the classification problem; (2) divides the annotated data set into training and validation data sets; (3) decides what machine-learning classifiers will be tested (e.g., support vector machines, k-nearest neighbor); (4) defines the time segment or data windows of the streaming data for applying the classifier (e.g., a classifier makes a decision for each 10-second window of data); (5) extracts features from the data windows that become the inputs for the machine-learning classifier (e.g., mean, variance, amplitude); and (6) tests the accuracy of the classifier (% agreement with annotated “truth”) with the training data set initially and then externally with the validation data set.[8,9,32,33] These steps are often conducted in an iterative process where

various classifiers, window sizes and data features are tested to find the best analytic strategy for the data.

Real-Time Data Analysis

Since many measures collected by mHealth can be obtained remotely in real time without the subject having to participate in traditional measurement visits, data analysis can be conducted more quickly, sometimes in real time. This can enable studies to be concluded earlier than planned when evidence of outcomes is obtained. Near real-time data analysis can be also used to control various versions of adaptive experimental approaches and designs. Additionally, the streaming data of mHealth can be used for real-time predictive modeling. This can help in the selective acquisition of measure and interventions, especially when these factors are not of equal expense. Real-time predictive modeling can provide evidence for adaptive selection and continuation within a broader study design.

Comprehensive Data Sets and Information Fusion

mHealth provides an opportunity to gather data from multiple sensors and modalities including divergent physiologic, behavioral, environmental, biological and self-reported factors that can be simultaneously linked to other indicators of social and environmental context. In addition, they can be linked to healthcare system and payer data at either the individual or population level. Combining or fusing these data, for example by using probabilistic techniques,[34] allows researchers to reduce measurement error and explore linkages among physiologic, behavioral, and environmental factors that may mediate or moderate treatment effects. Such ecologically rich data sets allow assessment of treatment effects under various real-world conditions. These comprehensive assessments may enhance the validity and reliability of the inferences and improve the statistical power of the assessment process.

In many cases, data fusion is useful to reduce the variability of the resulting estimates and to improve reliability. In some situations, however, data fusion can improve the validity of the estimates. For example, the interpretation of ambulatory electrocardiogram (ECG) data can be enhanced by data provided simultaneously by accelerometers: The ECG signal is expected to be strongly affected by physical disturbances due to activities such as running. Using the accelerometer information can therefore improve the interpretation of the raw ECG data and thereby reduce the probability of incorrect clinical interpretation that would lead to false alarms. What makes fusion challenging in practice is the fact that not all measurements can be made at the same spatial and temporal resolution. This presents multimodal and multiscale information fusion research with challenges that need to be addressed in future work.

Increasing the Number of Eligible Participants

mHealth can facilitate remote research recruitment and potentially reduce the frequency, and consequently, the burden of face-to-face interactions. The mobility of mHealth allows research to take place in a participant’s home, workplace, and community, rather than in trips to an academic research center. In addition to reduced travel, mHealth also has the potential to reduce burden by cutting down on required self-report, which can be augmented and, at times, replaced by non-invasive sensing. Finally, many mHealth tools can be scaled rapidly. Sensors and mobile phones provide researchers with untold research opportunities for both monitoring and intervening in real time. By increasing scalability, reducing burden, and spreading access to people beyond the reach of traditional health research, new populations may consider participating who would have never done so before. By broadening the participant pool for research studies, research can be not only more efficient, but also more generalizable.

Improved Adherence to Interventions and Assessment of Intervention Outcomes

mHealth methods can be used for real-time monitoring of treatment adherence and to discern factors that influence adherence behaviors. mHealth interventions may also be used to provide feedback and support to improve adherence. For instance, micro-payments can be provided at opportune moments to motivate protocol adherence,[35] and corrective actions can be initiated based on the detection of disturbances in adherence to the proper use of wearable technologies, such as loosening of wearable sensors.[9]

These same technologies can help reduce subject burden by substituting objectively measured assessments with those requiring participant engagement. Examples of these include the automatic detection of social interactions,[33] smoking,[8] and stress levels.[9] Finally, mobile technologies can be used to assess the effects of intervention dissemination, because the same feedback loops deployed to monitor intervention fidelity can be used to understand the impact of the intervention on treatment outcomes. The hypothesis that mHealth can improve evidence generation by offering new capabilities discussed in the preceding itself needs evidence to conclude that mHealth can indeed lead to better evidence.

Additional Issues for mHealth

Using Open Infrastructure and Data Standards

In addition to increasing research efficiency through design and technologic capabilities, mHealth technologies can enhance scientific efficiencies through the creation of modular platforms to share information and standardize and coordinate data collection. Based on the Internet ecosystem, the open platform specifies that substantial interfaces between the hardware and software components in the mHealth open system should be standardized and agreed on through collaboration among stakeholders. An open mHealth approach also specifies that interfaces should be published and available to the public. This model should reduce the need to create each project from “scratch” and should instead build on successful applications. For example, Open mHealth[36,37] is developing a repository of shared, open-source, interoperable modules for mHealth data interpretation and presentation for use by mHealth researchers and practitioners.

The open platforms also suggest that there is potential value in generating and promoting the use of common metrics and standards for mHealth research. These standards could include such things as “metatags” on sensor data to facilitate cross-study comparisons and the creation of databases to make data aggregation possible. Additionally, the use of common measures such as those available in the NIH PROMIS© program (Patient-Reported Outcomes Measurement Information System; www.NIHPromis.org); and the Neuroscience toolkit (www.nihtoolbox.org/Pages/default.aspx); and PhenX (www.phenx.org/) could help move the field forward by facilitating the same kinds of comparisons.

Privacy, Security, and Confidentiality

Several concerns exist about how privacy, security, and confidentiality in mHealth are handled, because these data can reveal highly personal information such as social interactions, location, emotion, and other potentially sensitive health conditions.[39] There has been a societal trend over the past few decades to accept the collection of person-level data for public good, such as the use of community-wide video surveillance for purposes of public safety. Nonetheless, the mHealth research community is now challenged to develop methods to preserve participant privacy and confidentiality while satisfying research needs.

Recent work suggests that this can be done using privacy-preserving protocols that address data confidentiality, authenticity, and integrity, as well as unlinking multiple data

transmissions.[39] This final component—unlinking—is critical because the loss of one data point/transmission is often not enough to create inferences that might compromise an individual's confidentiality, but multiple intercepted transmissions from a source can create a profile that becomes identifiable. This is especially the case when location data are included. Confidentiality, authenticity and integrity of the data can be accomplished through encryption, which transforms the data using a key known only to users on either end of the transmission. Unlinking requires a separate key that scrambles the linking features of each data packet, to which, again, only the sender and user have access.[39.40] These issues may be particularly important for research participants who have the greatest healthcare needs, such as older adults and low-income or disadvantaged populations.

mHealth also poses privacy challenges from people who are not enrolled in the research. Examples of this issue include the use of mobile cameras or microphones to collect data, but which also pick up sounds and images from nonparticipants. Similar to the issues raised at the participants' level, ways to address these issues are needed, not only at the level of study design, but also through the use of techniques that can extract information from raw data that abstracts the information while protecting privacy.

Conclusion

The capabilities inherent in mHealth constitute a new paradigm for evidence generation in health research, promising, perhaps more than any previous wave of innovations in health technologies, to help reduce the time from conception of interventions to their dissemination. Achieving this will necessitate addressing the many methodologic issues outlined above. Although these methodologic challenges present exciting new opportunities for scientific innovation, the marketplace and consumers are not waiting for scientific validation. This workshop endorsed the need for timely and increased efforts in mHealth research and for a new transdisciplinary scientific discipline incorporating medicine, engineering, psychology, public health, social science, and computer science. Training the next generation of mHealth scientists, a process recently begun via workshops sponsored by some the workshop sponsors (obssr.od.nih.gov/training_and_education/mhealth/index.aspx) will be essential if the health community is to realize the full measure of benefits from mHealth.

Acknowledgments

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the NIH or any other author-affiliated organizations.

VS has substantial financial interest in PinMed, Inc., which develops technologies and applications for mobile health. No other financial disclosures have been reported by the authors of this paper.

References

1. Krishna S, Boren SA, Bales EA. Healthcare via cell phones: A systematic review. Telemed J E Health. 2009; 15(3):231–240. [PubMed: 19382860]
2. Patrick K, Griswold W, Raab F, Intille SS. Health and the mobile phone. Am J Prev Med. 2008; 35(2):177–181. [PubMed: 18550322]
3. Riley WT, Rivera D, Atienza A, Nilsen W, Allison SM, Mermelstein R. Health Behavior Models in the Age of Mobile Interventions: Are Our Theories up to the Task? Transl Behav Med. 2010; 1:53–71. [PubMed: 21796270]
4. Quinn CC, Shardell M, Terrin M, Barr E, Ballew S, Gruber-Baldini AL, Cluster A. Randomized Trial of a Mobile Phone Personalized Behavioral Intervention for Blood Glucose Control. Diabetes Care. 2011; 34:1934–1942. [PubMed: 21788632]

5. Gupta S, Chang P, Anyigbo N, Sabharwal A. mobileSpiro: Accurate Mobile Spirometry for Self-Management of Asthma. Proceedings of ACM mHealthSys. 2011:Article 1.
6. Bexelius C, Lof M, Sandin S, Lagerros YT, Forsum E, Litton JE. Measures of physical activity using cell phones: validation using criterion methods. J Med Internet Res. 2010; 12(1):e2. [PubMed: 20118036]
7. Patrick K, Raab F, Adams M, et al. A text message-based intervention for weight loss: randomized controlled trial. J Med Internet Res. 2009; 11(1):1–9.
8. Ali A, Hossain SM, Hovsepian K, Rahman M, Kumar S. SmokeTrack: Automated Detection of Cigarette Smoking in the Mobile Environment from Respiration Measurements. Proceedings of ACM/IEEE Conference on Information Processing in Sensor Networks. 2012:269–280.
9. Plarre K, Raij AB, Hossain M, et al. Continuous inference of psychological stress from sensory measurements collected in the natural environment. Proceedings of ACM/IEEE Conference on Information Processing in Sensor Networks. 2011:97–108.
10. Burns MN, Begale M, Duffecy J, et al. Harnessing context sensing to develop a mobile intervention for depression. J Med Internet Res. 2011; 13(3):e55. [PubMed: 21840837]
11. D'eglise C, Suggs LS, Odermatt P. Short message service (SMS) applications for disease prevention in developing countries. J Med Internet Res. 2012; 14(1):e3. [PubMed: 22262730]
12. Abroms LC, Padmanabhan N, Thaweethai L, Phillips T. iPhone apps for smoking cessation: a content analysis. Am J Prev Med. 2011; 40(3):279–285. [PubMed: 21335258]
13. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat. 2007; 17(4):529–569. [PubMed: 17613641]
14. Piantadosi, S. Clinical trials: a methodologic perspective. New York, NY: Wiley; 2005.
15. Ioannidis JPA. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA. 1998; 279(4):281–6. [PubMed: 9450711]
16. West SG. Alternatives to randomized experiments. Curr Direct Psych Sci. 2009; 18(5):299–304.
17. Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin; 2002.
18. Kravitz RL, Duan N, Niedzinski EI, Hay MC, Subramanian SK, Weisner TS. What ever happened to n-of-1 trials? Insiders' perspectives and a look to the future. Milbank Q. 2008; 86(4):533–555. [PubMed: 19120979]
19. Brown C, Lilford R. The stepped wedge trial design: a systematic review. BMC Med Res Methodol. 2006; 6:54. [PubMed: 17092344]
20. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007; 28(2):182–191. [PubMed: 16829207]
21. Ratanawongsa N, Handley M, Quan J, et al. Quasi-experimental trial of diabetes self-management automated and real-timetelephonic support (SMARTSteps) in a Medicaid managed care plan: study protocol. BMC Health Services Res. 2012; 12(22)
22. Mohr, DC.; Duan, N. Continuous Evaluation of Evolving Interventions. mHealth Evidence Workshop; August 16, 2011; Washington DC.
23. Collins LM, Murphy SA, Strecher V. The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New Methods for More Potent e-Health Interventions. Am J Prev Med. 2007; 32(5S):S112–118. [PubMed: 17466815]
24. Lizotte DJ, Bowling M, Murphy SA. Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. Proceedings of Inter Conf Machine Learning (ICML). 2010:695–702.
25. Shortreed SM, Laber E, Lizotte DJ, Stroup TS, Pineau J, Murphy SA. Informing sequential clinical decision-making through reinforcement learning: an empirical study. Machine Learning. 2010; 84(1):109–136. [PubMed: 21799585]
26. Mitka M. Strategies sought for reducing cost, improving efficiency of clinical research. JAMA. 2011; 306(4):364–365. [PubMed: 21791679]
27. Shusterman V, Goldberg A, London B. Upsurge in T-wave alternans and nonalternating repolarization instability precedes spontaneous initiation of ventricular tachyarrhythmias in humans. Circulation. 2006; 113(25):2880–7. [PubMed: 16785339]

28. Hedeker D, Mermelstein RJ, Demirtas H. An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. Biometrics. 2008; 64:627–634. [PubMed: 17970819]
29. Steyvers, M.; Griffiths, T. Probabilistic Topic Models. In: McNamara, D.; Dennis, S.; Kintsch, W., editors. Handbook of Latent Semantic Analysis. Oxford, UK: Psychology Press; 2007.
30. Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc 18th International Conf on Machine Learning. 2001:282–289.
31. Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE. 1989; 77(2):257–286.
32. Preece SJ, Goulermas JY, Kenney LPJ, et al. Activity identification using body-mounted sensors--a review of classification techniques. Physiol Meas. 2009; 30(4):R1–33. [PubMed: 19342767]
33. Rahman M, Ali AA, Plarre K, al'Absi M, Ertin E, Kumar S. mConverse: Inferring Conversation Episodes from Respiratory Measurements Collected in the Field. Proceedings of ACM Wireless Health. 2011
34. Mitchell, HB. Multi-sensor Data Fusion – An Introduction. Berlin: Springer-Verlag; 2007.
35. Mustang M, Raij AB, Ganesan D, Kumar S, Shiffman S. Exploring micro-incentive strategies for participant compensation in high burden studies. Proceedings of ACM UbiComp. 2011:435–444.
36. Estrin, D.; Sim, I. Open mHealth. openmHealth.org/2010. openmHealth.org/
37. Estrin D, Sim I. Open mHealth architecture: an engine for health care innovation. Science. 2010a; 330(6005):759–760. [PubMed: 21051617]
38. Raij AB, Ghosh A, Kumar S, Srivastava MB. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. Proceedings of ACM CHI. 2011:11–20.
39. Mare, S.; Sorber, J.; Shin, M.; Cornelius, C.; Kotz, D. Adaptive security and privacy for mHealth sensing. USENIX Workshop on Health Security (HealthSec); August, 2011;

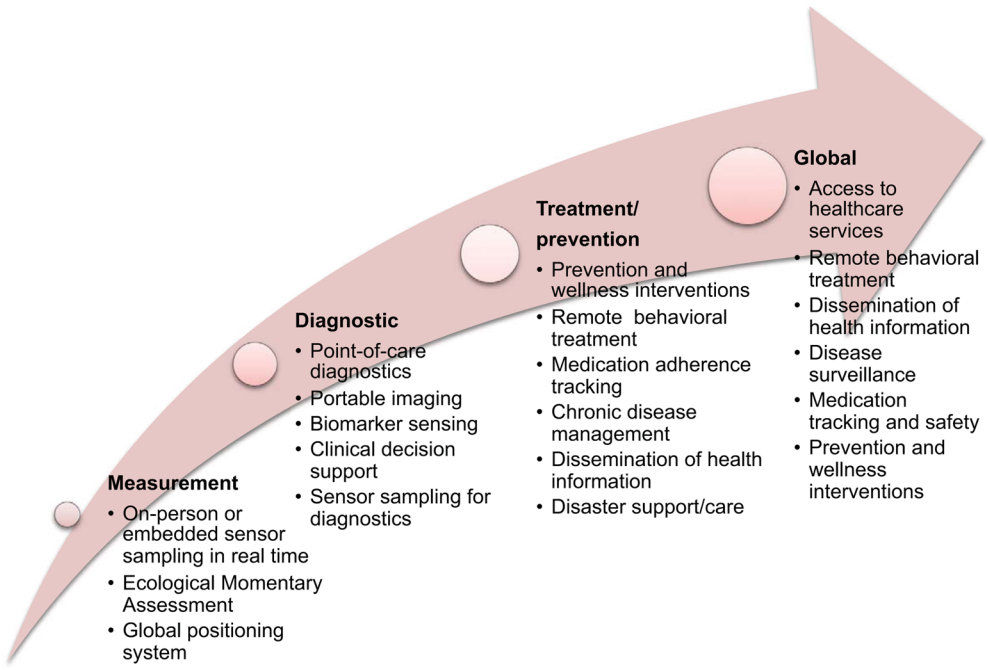

Figure 1.
Continuum of mHealth tools

Table 1

Participants in the mHealth Evidence Workshop at NIH, 2011, who also contributed to the current paper

Participants
Aman Bhandari, PhD
Sara Browne, MD
Tanzeem Choudhury, PhD
Linda M Collins, PhD
Deborah Estrin, PhD
Bob Evans
Paul Fontelo, MD
Craig Friderichs, MD
Deepak Ganesan, PhD
Margaret Handley, PhD, MPH
Bethany Hedt, PhD Susan Horn, PhD
Robert Kaplan, PhD
Thomas Kirchner, PhD
Joseph Kvedar, MD
Marc Lara, PharmD, MBA
Garrett Mehl, PhD
Barbara Mittleman, MD
Inbal Nahum-Shani, PhD
Greg Norman, PhD
Bakul Patel, MS, MBA
Michelle Rodrigues, MBA
Ida Sim, MD, PhD
Carrie Varoquiers, MPH
Tisha Wiley, PhD
Hui-Lee Wong, PhD

Table 2

Reliability and validity in mHealth

	Construct	Challenge for mHealth	Example
Reliability	Reliability refers to the consistency of a measure. A measure is said to have a high reliability if it produces consistent results under consistent conditions.		
Test–retest	The degree to which assessment values are consistent when repeated	May be a challenge when the goal is to capture temporal variability	Self-reported mood collected daily via Ecological Momentary Assessment
Inter-method reliability	The degree of agreement among various assessment methods	No challenges noted and actually may be very appropriate for mHealth	Two different types of accelerometers worn on the same wrist
Validity	Validity is considered to be the degree to which an assessment measures what it claims to measure.		
Concurrent validity	Different measures of the same phenomena should produce similar results	No challenges noted	Correlation in the measures of conversation via respiration or microphone
Convergent validity	Degree of agreement between a new assessment method and a gold standard	Many mHealth assessments are new ways to assess constructs with no gold standard (or “ground truth”)	Wireless plethysmography of respiration patterns validated against a clinically accepted stationary unit
Divergent validity	Degree to which the new measure diverges from measures of other phenomena	No challenges noted	Wireless measures are not correlated with height
Predictive validity	How well a future outcome can be predicted from the measure	No challenges noted, and intensive data collection may enhance predictive ability	Myocardial infarction predicted by mobile electrocardiogram

Table 3

Potential research designs to evaluate the efficacy and effectiveness of mHealth interventions.

	Key Points	Additional Considerations
Treatment Development Stage	The following quasi-experimental designs are useful during treatment development. They are also important for studies in which randomization is not possible.	
Pre–post test designs[17]	Target measures that are collected before the intervention begins (pretest) serve as the baseline. Post-test measures are used to estimate change due to the intervention. Causality cannot be determined because there is no control or comparison group, and potential confounding variables limit interpretation of effects.	Can be used with continuous measurements to examine changes in the trend of target behavior over time. Measures of potential confounding variables help estimate change due to the intervention and assess variation introduced by other factors.
n-of-1 design[18]	Multiple cross-over, single-subject experimental design that is an experimental variant of the pre–post text design. It reduces bias and accounts for random variation over time through repeated, short-term comparisons of treatments, where each treatment may be an active intervention or placebo. The patient receives treatments in a systematic or random sequence.	*n*-of-1 trials are applicable to chronic symptomatic conditions and to treatments characterized by rapid onset of action and quick washout. Obviously, this design is only feasible when ethically appropriate.
Interrupted Time-Series Design	Pre–post test design where large numbers of repeated measurements are collected before and after the treatment. The premise is that administration of the treatment should produce an interruption to the pre-treatment time series. The interruption can be found along any of three dimensions: form of the effect (the level, slope, variance, cyclicity); permanence of the effect (continuous or discontinuous); and immediacy of the effect.	This design is especially suited to mHealth, in which multiple measurements are common.
Mature Intervention Testing	When interventions have been developed that are feasible and usable and have quasi-experimental or pilot data supporting their efficacy, larger randomized trials are appropriate.	
RCT	The participants, groups or locations are randomly assigned to treatment or control group.	This is the most common trial design for testing causality in health research.[14]
Regression discontinuity design	Participants are assigned to treatment or control based on whether they fall above or below a criteria cutoff score. The assignment variable may be any variable measured before treatment. The design is similar to interrupted time series, but differs in that the effect or interruption occurs not at a specific time, but rather at a cutoff score in regression discontinuity.	The design is most powerful when the cutoff is placed at the mean of the assignment variable since the analysis focuses on the subjects close to the cutoff score. Since only a fraction of the participants are used for the analysis, this technique requires more participants in otder to equal the power of a randomized experiment.
Stepped-wedge design[19–21]	This design operates as a series of waiting lists and randomizes the order in which groups, locations or even populations receive the intervention. The intervention group can be compared with both their pretest measures and with measures from other subjects who have not yet received the treatment, who form an independent and homogeneous control group at each time point.	In this design, all participants are told that they will receive intervention, which ensures participants are not denied intervention. Stepped-wedge design is appropriate if the intervention is going to be implemented with all individuals (or at all sites) and if it is not feasible to scale all at once.