Opportunities and Challenges in the Collection and Analysis of Digital Phenotyping Data

Running Title: Opportunities and Challenges in Digital Phenotyping Data

Jukka-Pekka Onnela, D.Sc.
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Harvard University

Email: onnela@hsph.harvard.edu

Abstract

The broad adoption and use of smartphones has led to fundamentally new opportunities for capturing social, behavioral, and cognitive phenotypes in free-living settings, outside of research laboratories and clinics. Predicated on the use of existing personal devices rather than introduction of additional instrumentation, smartphone-based digital phenotyping presents us with several opportunities and challenges in data collection and data analysis. These two aspects are strongly coupled, because decisions about what data to collect and how to collect it constrain what statistical analyses can be carried out, now and years later, and therefore ultimately determine what scientific, clinical, and public health questions may be asked and answered. Digital phenotyping combines the excitement of fast paced technologies, smartphones, cloud computing and machine learning, with deep mathematical and statistical questions, and it does this in the service of a better understanding our own behavior in ways that are objective, scalable, and reproducible. We will discuss some fundamental aspects of collection and analysis of digital phenotyping data, which takes us on a brief tour of several important scientific and technological concepts, from the open source paradigm to computational complexity, with some unexpected insights provided by fields as varied as zoology and quantum mechanics.

Introduction

It has been estimated that by the end of 2022, there will be 6.3 billion smartphone subscriptions globally (1). We now depend on our smartphones for organizing and navigating our daily lives. Just as walking on a beach leaves a trail of footprints in the sand, the use of these devices generates, as a byproduct, digital trails of social, behavioral, and cognitive footprints. Given that these trails reflect the lived experiences of people in their natural environments, it should be possible to leverage them to develop precise and temporally dynamic disease phenotypes and markers to diagnose, monitor, and treat illnesses (2, 3). Although the argument advanced here rests on precise measurement, in the words of the great 19th century theoretical physicist James Clerk Maxwell, the real reward for the "labor of careful measurement" is not just greater accuracy but the "development of new scientific ideas" and the "discovery of new fields of research" (4). We call this field of research *digital phenotyping* (2, 3), which we have previously defined as the "moment-by-moment quantification of the individual-level human phenotype *in situ* using data from personal digital devices, in particular smartphones." This field of digital phenotyping is therefore part of a larger field of research, known as deep phenotyping (5).

This review addresses the scientific problem sometimes known as the *phenotyping problem*. Over the past 20 years, many researchers have advocated a more substantial role for large-scale phenotyping as a route to advances in the biomedical sciences. Arguably we are held back by the phenotyping problem, our inability to precisely specify phenotypes, the observed manifestations of our genomes within our lived environments, in the individuals we study and treat. With over 10,000 named diseases, precision medicine and precision public health require a better understanding of phenotypes for analysis of phenomic data per se, as well as for the joint analyses of phenotypes and genotypes. Of the many different phenotypes, social, behavioral and cognitive phenotypes have traditionally been challenging for phenomics because of their temporal nature, context dependence, and the lack of tools for measuring them objectively in naturalistic settings. Pen-and-paper surveys are still widely used to solicit these phenotypes although they suffer from well-documented biases, including the tendency of individuals to reconstruct, rather than recall, their past. The phenotyping

problem is especially severe in psychiatry and neurology, where precise markers are desperately needed, but individuals suffering from these conditions may not be able to provide accurate self-reports.

The ubiquity of smartphones offers a scalable solution to the phenotyping problem. We have previously defined and operationalized the concept of *digital phenotyping* as the "moment-by-moment quantification of the individual-level human phenotype *in situ* using data from personal digital devices, in particular smartphones" (2, 3). Compared to existing approaches, smartphone-based digital phenotyping, especially when using passively and unobtrusively collected smartphone sensor and log data, enables large scale, long-term phenotyping in naturalistic settings with minimal subject attrition. Our early work, started in 2005, made use of anonymized cell phone call detail records (CDRs) to learn about the structure of large-scale social and communication networks (6). Given that most diseases impact behavior, cognition, and social functioning, it is perhaps not surprising that smartphone-based digital phenotyping has turned out to be broadly applicable. However, what has turned out to be more challenging than expected is how to make sense of millions of data points and how to turn them into insights.

In this review, we focus on opportunities and challenges in the collection and analysis of digital phenotyping data. This is not intended to be a technical review, but instead we focus on the fundamentals of collection and analysis of digital phenotyping data, with some examples from the literature. The references are meant to be illustrative rather than exhaustive. This is an incipient field that appears to be growing very quickly, but these fundamentals are likely to remain valid for years to come.

What is Digital Phenotyping?

Here we examine our definition of digital phenotyping as the "moment-by-moment quantification of the individual-level human phenotype *in situ* using data from personal digital devices, in particular smartphones." Although definitions may evolve, in the days of precision health and precision medicine it is imperative to provide precise definitions for terms. A discussion of the definition also highlights opportunities and challenges in data collection and data analysis.

The "moment-by-moment quantification" part is intended to emphasize collection and analysis of data that is generated continuously rather than at discrete points in time, often called waves in longitudinal studies. In most longitudinal studies, the goal is to assess each individual on an identical set of occasions, for example, at two points in time corresponding to the outcome of interest before and after an intervention. If data is collected at n points over time, in the analysis stage we will typically consider n observations over time. In contrast, when data are collected continuously, sometimes referred to as *temporally dense data*, the number of data points may be hundreds of thousands per day per subject. These data are usually first summarized, say, in 24-hour windows, and the data summaries are used in subsequent data analyses. Occasionally the terms continuous and real-time are used interchangeably, but they refer to two different concepts: continuous means that data collection happens in continuous time rather than at discrete points in time, whereas real-time refers to data processing and data analysis, the ability for a computational procedure to generate its output very shortly after having received its input. In our problem setting, real-time processing requires continuous data collection, but continuous data collection does not imply real-time processing.

The "individual-level human phenotype" part emphasizes the fact that the data are collected at the level of individuals rather than at group level, but it also accentuates the fact that many data analyses focus on within-person changes over time (7). This is in part necessary because the way people use their smartphones may be highly idiosyncratic, which makes comparisons across individuals difficult whereas within person comparisons, where every individual acts as their own control, remain valid. Group-level data are useful for example for studying the flow of individuals across cities, such as monitoring compliance with non-pharmacological interventions in the midst of a pandemic (8), but cannot be used for making inferences about individuals without committing ecological fallacy, i.e., making inferences about individuals from inferences about the group to which those individuals belong (9).

The "*in situ*" part highlights the fact that the data collection occurs in naturalistic or free-living settings. This part of the definition also implicitly references the use of passive data, data collected from smartphone sensors and logs without any burden on the subject, which is important for capturing "real world" behavior. Asking subjects to take frequent surveys or to complete assessments on the phone is not naturalistic. A good example of the importance of passive data is the Apple Asthma study (10). In the study, the participants were asked to complete intake surveys, daily surveys on asthma and medication adherence, and weekly surveys on healthcare utilization and quality of life. Out of 40,683 users who downloaded the study application in the U.S., 7,593 were enrolled after confirming eligibility. There were 6,470 "Baseline Users" who responded to at least one survey; 2,317 "Robust Users" who completed 5 daily or weekly surveys; and 175 "Milestone Users" who completed the 6-month survey. This attrition, from 7,593 to 175 subjects, translates into a loss of 97.7% of the cohort over a 6-month period. The authors of the study conclude that the study design is a good match for studies with a "hypothesis that can be answered in a short time period (1–3 weeks)." However, by relying more heavily on passive data collection and using a much lighter active data component, our group and others have followed subjects for several months or years. For example, in a study at McLean Hospital we have collected data from a cohort of bipolar patients continuously for 4.5 years (11). This emphasizes the importance of having a minimally burdensome approach to data collection, which is especially important for chronic conditions.

The "data from personal digital devices" part highlights the importance of using devices people already own and use rather than introducing additional instrumentation. In the same manner as asking subjects to use their own devices in unnatural ways, such as for taking frequent surveys, is likely to lead to attrition, the same can be said about introducing an additional device, whether it be a loaner phone or a wearable. Individuals do not appear to use loaner phones the same way they use their own phones, and in addition the use of loaner devices creates logistical challenges. A recent study implemented a 30-day loaner iPhone and smartwatch recirculation program as part of an mHealth intervention to improve recovery and access to therapy following acute myocardial infarction (12). Of participants enrolled with a loaner phone, 72% (66/92) returned the phone and 70% (64/92) returned the watch. The study reported a 61% cost saving using loaner devices compared to purchasing an iPhone for each participant who did not already own one. While optimizing loaner returns could lead to further cost savings, the use of loaner devices appears to be costly given the short study duration. The "digital device" part is not intended literally as a contrast to analog devices, just like machine learning is not a contrast to human learning, but instead refers more broadly to digital consumer electronics devices that can be used readily to collect data and are programmable.

The final part of the definition focuses on smartphones. If smartphones are to be part of the solution of the phenotyping problem, they should be broadly available to individuals across demographic factors, in particular sex, age, race/ethnicity, and socioeconomic status. Although a digital divide does exist in this area, empirical evidence suggests that it is rapidly narrowing. Smartphone ownership among U.S. adults increased from 35% in 2011 to 80% in 2018. Smartphone ownership is especially high among young adults: 96% of Americans ages 18-29 own a smartphone. A recent study implemented in the UK surveyed 2,167 5-16-year-olds, finding that by age 11, 90% had their own device, and phone ownership was "almost universal" once children were in secondary school(13, 14). This suggests that smartphone-based digital phenotyping may be especially well suited to studying adolescents given that half of all lifetime cases of mental illness have an onset of age 14. Globally, 6.3 billion smartphone subscriptions are expected by 2022. While mobile phone and smartphone ownership is higher in the general population than in people with serious illness, this is also beginning to change. And while men and women appear to use smartphones somewhat differently (e.g., mean daily use 154 vs. 167 minutes) (1), our recent analysis of data quality from various cohorts demonstrates there is no difference in the quality of the collected data from men and women (15). Recruitment and retention of underrepresented minorities in research has traditionally been difficult and expensive, but with ever increasing smartphone ownership this disparity will hopefully be mitigated, although it is still a matter of debate (16).

There are two fields that are closely related to digital phenotyping, remote patient monitoring and mobile phone sensing. *Remote patient monitoring* (RPM) refers to the use of a non-invasive, wearable devices that automatically transmit data to some back-end system or smartphone application for patient self-monitoring and/or health provider assessment and clinical decision-making. A recent RPM meta-analysis identified 27 randomized controlled trials (RCTs) that focused on a range of clinical outcomes (17). The studies had an average duration of 7.8 months and an average sample size of 239 patients (range 40–1437). The RPM devices employed in these studies included blood pressure monitors, ambulatory electrocardiograms, cardiac event recorders, positive airway pressure machines, electronic weight scales, physical activity trackers and accelerometers, spirometers, and pulse oximeters. For example, a study of stroke patients used ankle accelerometers and two types of feedback to assess the difference in average daily time spent walking between treatments (18). Some studies also incorporated the smartphone in the loop, such as when using smartphones to provide self-care messages after each blood pressure reading (19). Most patient remote monitoring applications introduce additional instrumentation to the lives of patients. Studies have found that most devices result in only short-term changes in behavior and motivation (20), and activity trackers have been found to change behavior for only approximately three to six months (21). According to an often-cited study (n=6223), most individuals who purchased a wearable device stopped using it and, of these, one-third did so before six months (22). Reasons for abandoning wearables include users not finding them useful or the devices breaking (23). RPM appears to be most useful in settings where there are opportunities to use the data to change clinical care in settings that last at most for a few months, for example, as part of rehabilitation.

Mobile phone sensing refers to the use of various types of sensor data to enable new forms for social networking, sensor augmented gaming, virtual reality, and smart transportation systems (24). The umbrella term that is often used is *urban sensing* because most mobile phone sensing systems are being deployed and used in urban areas. Urban sensing is usually divided into *participatory sensing* and *opportunistic sensing*; in the former the participant is directly involved in the sensing action (e.g., by tagging locations), whereas in the latter the user is not aware of sensing or involved in decision making. *People-centric urban sensing* focuses on different aspects of a person and her social setting, for example, where she is and what she is doing. *Personal opportunistic mobile phone*

sensing has many applications in health and wellness ranging from fall detection systems to cardiovascular disease management systems. Well known platforms in this area include EmotionSense, which was an Android application part of a research project that ran between 2011-2016 (25), intended for sensing emotions and activities of individuals, and Darwin (26), which is intended for reasoning about human behavior and context. These and other similar platforms are now increasingly used to study health and wellness, although a large majority of studies appear to make use of non-clinical cohorts. For a listing mobile phone sensing software, see the Wikipedia page (27). Sensing is clearly a much broader field than digital phenotyping and has various goals. Although both use the smartphone for data collection, this similarity may be deceiving. As an analog, ornithologists use binoculars and telescopes to study birds, but the same instruments are used by astronomers to study the skies. Digital phenotyping has as its core one mission: to provide more granular and more precise phenotypes. In contrast, in mobile phone sensing, the goal is usually to solve a specific problem, such as transportation, which is what companies like Uber do.

Challenges in Data Collection

Data collection happens through a smartphone application the research subject downloads and installs on her personal smartphone after having first consented to participate in the study. Because of the personal nature of the collected data, it is of paramount importance that subjects understand what data is collected and for what purpose. A subject can leave the study at any point by selecting the appropriate option within the study application. Deleting the application will stop any ongoing data collection. All smartphone data collected in digital phenotyping can be divided into two main categories: active and passive. *Active data* refers to anything that requires the user to actively participate in a data collection activity as a result of her participation in a study. Examples of active data include taking surveys, contributing audio diary entries, or using the phone to carry out cognitive assessments. *Passive data* generally refers to data that is generated or can be generated by the device passively (from the point of view of the user) and they pose no burden on the participant, thus constituting objective measurement of different aspects of social, behavioral, and cognitive functioning.

Active data is always associated with at least some level of subject burden, and therefore introduces the challenge of how to keep subjects engaged over long time periods. Use of financial incentives is a common approach to boost engagement in small studies over short time periods, but it constitutes an intervention that may distort the quality of the data and is not economically feasible in large cohorts. Small financial incentives may work well with college students looking to earn some extra income but might be less successful when used with financially secure individuals. In some settings, the subjects have strong incentives to respond a certain way. For example, in professional sports athletes may earn more in a single game than a professor earns in a year, and this clearly introduces a strong bias to respond a certain way on post-concussion symptoms. Other common approaches for improving engagement is to provide feedback or use gamification. But gamification is notoriously difficult to do well, even when the sole goal is entertainment, and a large majority of commercial games fail, the average "good" indie game making an estimated $25,000 in its first year on sale (28).

Passive data originates from smartphone sensors (such as GPS and accelerometer) and smartphone logs (such as communication logs and screen activity logs). The reason this distinction between sensors and logs matters is that a smartphone application has to request access to a sensor for that data stream to become available,

and data collection starts after the application is running. There is no way to collect sensor data retroactively. The phone may have other applications running that have collected data, or the phone's operating system may have collected data for its own purposes, but these data are not available to the study application. In contrast, some smartphone logs are available and accessible for times that precede installation, such as Android communication logs, capturing metadata for phone calls and text messages

It is worth stressing that passive data being objective does not imply it being perfect; surveys certainly are not perfect either. But for constructs that can be assessed both using subjective surveys and objective passive data, it is hard to come up with a convincing argument in favor of surveys. Most people are aware of their own weight and height and could easily report them, yet we are all asked to step on a scale and have our height measured when we visit a healthcare provider. Neither measurement is perfect, but they are objective and more accurate than self-reported equivalents. An important caveat is that while some constructs, such as psychomotor agitation or sleep, can be more readily measured using passive data, for more complex constructs in psychiatry, such as anhedonia, we are still trying to determine the best active and passive measures. Also, passive data collection certainly has its challenges: it is difficult to implement research grade passive data collection; some sensors generate very large volumes of data and others cause significant battery drain; analysis of raw passive data is difficult; and collection of certain types of data may interfere with the phenotype of interest, for instance, keyboard-based data collection may affect user experience and their typical speed or usage. Here we focus on the first challenge; discussion of the second and third is presented in Challenges in Data Analysis.

When discussing passive data, it is important to distinguish between collecting *raw data*, such as obtaining samples of device acceleration along three orthogonal axes multiple times per second, vs. collecting *predefined data summaries*, such as daily step count, which may use accelerometer data as an ingredient when generating that summary. Another important distinction is between what is possible when the application is running in the *foreground*, meaning that the application is shown on the screen and the user can interact with it, vs. when it's running in the *background*, when the user is not using the application.

The introduction of software development kits (SDKs) for smartphones, such as Apple's ResearchKit and Google's ResearchStack, has facilitated the writing of software for these devices, but it is important to appreciate their shortcomings. Use of pre-packaged software limits what type of data can be collected, which then limits what types of analyses can be performed. The problem with this approach is that what is available and convenient can begin to shape the research agenda of investigators, which is the reverse of how science usually works: formulate the question first; then determine the study design, including what data needs to be collected and how it needs to be sampled; next perform data analyses, which are always conditional (depend on) study design; and finally interpret the results, which is conditional on all previous steps. Perhaps the greatest potential of smartphones as a research tool, the fact that we carry them on our person, is substantially reduced if the application can only collect data when it is running on the foreground. This means that some applications, although they do collect passive data, only do so for the minute or two when the person is using the application. This brings many of the limitations of active data collection to passive data collection; although the data are objective *when* collected, if the user is not using the application, no data will be collected. For example, Apple's ResearchKit has been extremely popular and since its introduction in 2015 has been used to power countless of applications, but one of its limitations is that it does not support background sensor data collection (29). Although it implements an impressive range of various tasks, such as a timed walk

task and a gait and balance task (30), a person would have to use it fairly consistently to see changes in these behaviors as part of disease progression or treatment response, which could take months or years. Other SDKs, such as HealthKit, do support background sensor data collection but only in a limited manner. At the time of writing, it is possible to use a HealthKit function to record up to 12 hours of accelerometer data in the background; in addition, this capability does not appear to exist for other sensors (31). Continuous background data collection usually requires custom code. Other challenges with SDKs are that data generated by different SDKs are not comparable, so at the very least one would need to stratify the cohort to Android and iOS users, but this introduces a significant confounder given that iOS users have on average much higher annual incomes than Android users, approximately $53,000 vs. $37,000 (32). Some of the algorithms that are used to generate summaries by these SDKs are proprietary and are subject to change. As an analogue to the weight scale example, this would be the equivalent of a person monitoring their weight when the scale is being tweaked occasionally without them knowing about it.

In the past decade, there has been increased attention paid to the problem of transparency and reproducibility of science. One study found that only 6% of the sampled medical studies were completely reproducible (33), and a survey published in the journal Nature in 2012 found that 47 out of 53 medical research papers on cancer were irreproducible (34). A comparison of five analyses of reproducibility in preclinical research reported prevalence of irreproducibility to vary from 51% to 89% (35). Use of proprietary algorithms and proprietary metrics is problematic from the point of view of transparency and reproducibility of research, and many smartphone-based studies do not disclose enough details to even allow for attempts at replication. Algorithmic fairness, accountability, and transparency has recently emerged as its own interdisciplinary research area. Some of the relevant considerations have to do with, for example, generalizability. If a pre-packaged closed algorithm, say for counting steps, has been trained using data from young men, it may be accurate when used by users of this group, but it will likely perform less well on elderly women, and may perform poorly on elderly women with a neurodegenerative disorder. Designing complex algorithms is difficult, but it is the closed nature of some of these algorithms that makes it difficult to even detect a problem, let alone fix it. Another problem with the use of pre-packaged summaries is that new summaries cannot be implemented post data collection, and because of changes in underlying closed algorithms, data cannot be readily pooled across studies for later re-analyses.

Collection of private personal data from smartphone presents many challenges to privacy and data security, and here we will only touch the surface of this important and complex topic. We address two points, data encryption and anonymization. The starting point is that collected data must be encrypted at all times. For example, on Beiwe, the high-throughput research platform for smartphone-based digital phenotyping developed by our group, all data are encrypted while stored on the phone awaiting upload and while in transit, and they are re-encrypted for storage on the study server while at rest. More specifically, during study registration the platform provides the smartphone app with the public half of a 2048-bit RSA encryption key. With this key the device can encrypt data, but only the server, which has the private key, can decrypt it. Thus, the Beiwe application cannot read its own data that it stores temporarily, and there is therefore no way for a user (or somebody else) to export the data. The RSA key is used to encrypt a symmetric Advanced Encryption Standard (AES) keys for bulk encryption. These keys are generated as needed by the app and must be decrypted by the study server before data recovery. Data received by the cloud server is re-encrypted with the study master key provided and then stored on the cloud.

Some of the data that are collected in smartphone-based digital phenotyping studies contain identifiers, such as phone numbers that are part of communication logs on Android devices. The phone numbers and other similar identifiers, such as MAC addresses of Wi-Fi networks and Bluetooth devices, need to be anonymized. Again, using the Beiwe platform as an example, every phone generates a unique cryptographic code, called a salt, during the Beiwe registration process, and then uses the salt to encrypt phone numbers and other similar identifiers. The salt never gets uploaded to the server and is known only to the phone for this purpose. Using the industry standard SHA-256 (Secure Hash Algorithm) and PBKDF2 (Password-Based Key Derivation Function 2) algorithms, an identifier (e.g., phone number) is transformed into an 88-character anonymized string that can be used in data analysis. Note that for any given identifier, such as a phone number, by design always corresponds to the same 88-character string, which makes it possible to distinguish between two phone calls coming from one person vs. two individuals each placing one call. This is important because the structure and dynamics of communication networks can be used to assess social and cognitive functioning passively in addition to smartphone GPS and voice data that are currently used for assessing cognition (36). A recent study demonstrated the connection between social networks and functional recovery of patients with ischemic stroke (37): using a quantitative social network assessment tool (rather than smartphone-based communication data), the researchers demonstrated that an average patient's network over 6 months contracted by 1.25 people. Anonymization can also be implemented for other data streams, such as GPS, which can accurately resolve the location of a person's home. On the Beiwe platform, investigators can choose to use the "noisy GPS" data stream which incorporates additive noise to the coordinates. Naturally, all summary statistics derived from noisy data will themselves be noisy, but in some cases this is the preferred tradeoff between privacy and accuracy.

Another challenge with digital phenotyping data is data sharing, which is critical for transparency and reproducibility of research. For example, the National Institute of Mental Health Data Archive (NDA) makes available human subjects data collected from hundreds of research projects across many scientific domains. NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science and discovery. However, there are unresolved technical problems on how to best share smartphone data. One possible solution might be to share intermediate abstract data types rather than the raw data itself. For example, GPS data are often converted into an intermediate data type, a sequence of flights (periods of linear movement) and pauses (periods of non-movement). Sharing of these types of data might be an intermediate solution while the scientific community comes to consensus on best data sharing practices.

Opportunities in Data Collection

The Precision Medicine Initiative (PMI) was intended to approach health from a broad perspective, taking into account individual variability in genes, environment, and lifestyle. It focuses on disease onset and progression, treatment response, and health outcomes through the more precise measurement of molecular, environmental, and behavioral factors that contribute to health and disease. As the PMI Working Group noted in their report (38), these developments are now possible because of "advances in genomic technologies, data collection and storage, computational analysis, and mobile health applications." The authors then continued with the following: "Data from sensors and software applications can enrich self-reported data on lifestyle and environment, giving researchers a clear view into these factors that have previously been difficult to capture with accuracy." It is notable that of the three factors, smartphones can directly capture behavioral factors and can indirectly capture environmental factors; for example, starting

from time-varying measurements of, say, pollution in specific areas, one can use timestamped location data to estimate how much time a person spends in different areas, and then use these time estimates as weights to calculate a personalized, individual-level exposure to pollution. Many other similar applications are possible. Smartphone-based digital phenotyping clearly fits in with the precision medicine paradigm.

Digital phenotyping is also consistent with the goals of U.S. National Institute of Mental Health (NIMH), which in 2018 published a notice (39) to highlight its interest in receiving grant applications that utilize digital health technology to advance assessment, detection, prevention, treatment, and delivery of services for mental health conditions. A workgroup explored opportunities and challenges of digital health technology relevant to the NIMH mission, and identified three areas as a high research priority: (1) assessment, (2) intervention refinement and testing, (3) and service interventions and service delivery (40). Digital phenotyping is clearly responsive to the first area, assessment, which identifies "technology-assisted … collection of behavior in natural environments … to create digital/behavioral phenotypes" as a high priority for the NIMH. Digital phenotyping is also compatible with the Research Domain Criteria (RDoC) research framework for studying mental disorders. As defined by the NIMH, the framework consists of a matrix where the rows represent dimensions of function (Domains and Constructs) and the columns represent areas for study (Units of Analysis) (41). The five domains of the RDoC matrix are negative valence systems (responsible for responses to adverse situations), positive valence systems (responsible to positive situations), cognitive systems (responsible for cognitive processes), systems for social processes (mediating responses to interpersonal settings), and arousal / regulatory systems (responsible for generating activation of neural systems). Smartphones offer novel tools to capture several units of analysis of the NIMH's RDoC framework, making it possible to collect new data streams that were previously difficult or nearly impossible to capture.

Experience sampling methods (ESM) and ecological momentary assessment (EMA) are survey methodologies that can offer insight into daily life experiences, including symptoms of mental disorders. Retrospective clinician-administered and self-report questionnaires are the gold standard in human psychopharmacology. Unfortunately, retrospective measures are subject to memory distortions, and may more likely reveal how patients reconstruct the past, not how they experienced it (42), and current mood is likely to influence the type of information that is recalled (43). Retrospective recall of average levels of mood or symptoms might be also more difficult than considering the present moment, particularly for individuals with psychiatric diagnoses (44). For example, one study found that retrospective reports of extreme mood changes, over the previous month and even over the preceding week, were largely unrelated to reports obtained *in situ* (45). As ESM/EMA questions pertain to the present moment or to a recent interval, memory biases are expected to be minimal, and the methodology taps into processes and experiences as they occur in real life. Early ESM/EMA studies used paper questionnaires, but computerized versions on personal digital assistants (PDAs) and smartphones have also become available (46). There are reviews on ESM/EMA studies on major depressive disorder (47), psychotic disorders (48), substance use disorders (49), anxiety disorders (50), and eating disorders (51). A fairly recent review cuts across diagnostic categories by including 18 studies that applied ESM/EMA to study the effects of medication on patients with major depressive disorder, substance use disorder, attention-deficit hyperactivity disorder, psychotic disorder, and anxiety disorder (46). Of these 18 studies, 11 used paper, 5 a personal digital assistant (PDA), and only 2 used smartphones. ESM/EMA based on paper may suffer from poor compliance because subjects may not complete the questionnaires as instructed. The introduction of a PDA, which studies have introduced as an additional device, suffers from all the usual difficulties when a person is asked to carry on their person an additional device. In addition to being

able to timestamp responses and thus quantify whether subjects took the surveys when pinged, additional information can be recorded, such as the location where the survey was taken. Smartphones appear to be the ideal instrument for ESM/EMA, with additional insights possible when coupled with other types of smartphone data.

An important opportunity in this space is the availability of open source research platforms. We use the term platform here to draw a contrast to standalone smartphone applications that serve specific use cases. For example, the iOS application used in the asthma study discussed earlier is an application but not a platform. A platform should support ideally both front-ends, smartphones running Android and iOS operating systems, should enable customization in active and passive data collection, and should support study monitoring and data analysis activities, and as such typically consist of separate front-end and back-end components. Stated differently, a standalone application here is something that a person can use for a specific use case, whereas a platform is a system or collection of multiple pieces of software to support the running of research studies. Beiwe is an example of an open source research platform for smartphone-based high-throughput digital phenotyping. The native Android and iOS smartphone applications that constitute the front-end of the configurable platform collect various types of active data, such as surveys and audio diary entries, and passive data, such as GPS and accelerometer data in their raw (unprocessed) form, and anonymized phone call and text messages logs. The Beiwe back-end, which is based on Amazon Web Services (AWS) cloud computing infrastructure–making it both scalable and globally accessible–collects, stores, and analyzes the collected data. These data enable us to study behavioral patterns, social interactions, physical mobility, gross motor activity, and speech production among other phenotypes. The platform has been developed and maintained by professional software engineers; both front-end and back-end code are open source and available under the permissive 3-clause BSD license.

There appears to be some confusion about what open source software means, and since we identify it as a major opportunity, it is worth clarifying this terminology. Part of the confusion likely arises from the fact that the word "free" in English has at least two different connotations. The commonly used examples in this context are that software can be free, as in free beer, but software can also be free, as in freedom of speech, which means that the user is free to modify and distribute that software. For example, Adobe Reader used for reading PDF documents is what is often called *freeware*: it is free in the former sense but not in the latter; the source code is not available, and the license prohibits you from trying to reverse engineer the software. The web browser Firefox is what is often called *free software*: it is free in both senses. Beiwe is *open source*, which means what one would assume it to mean, which is that its source code is publicly available, making it free in the first sense; because the source code is in the public domain under a permissive license, it is also free in the second sense. However, just because software is free, it does not mean that *running* that software incurs no cost to the user. Just like a person using a home computer needs to pay for the computer and its maintenance (often by taking time installing updates and doing similar tasks), the cost of running software or a computing platform is distinct from the cost associated with purchase or license to use software.

Readers unfamiliar with open source software may wonder whether open source software is secure, whereas people familiar with the paradigm usually assert that open source software is *more* secure than closed source software for the reason that the code is available for inspection and study by anyone, and therefore software errors and security vulnerabilities are more likely to be discovered. The Linux operating system is one of the most successful examples of open source software, and according to one estimate, 96.3% of the world's top

one million servers run on Linux (52). There is obviously variability across different projects, whether open source or closed source, but a rare empirical analysis comparing published vulnerabilities concluded that the mean time between vulnerability disclosures was lower for open source software in half the cases and comparable in the other half, and there were no significant differences in the severity of vulnerabilities between open source and closed source software (53).

Challenges in Data Analysis

The main difficulty in the analysis of smartphone data is a direct consequence of the main opportunity in the collection of such data: smartphones are personal devices that are used frequently by many people, over long periods of time, in a myriad of ways. Much before the arrival of smartphones, over two decades ago, zoologists started using accelerometers to study animal behavior (54). Field biologists can rarely observe animals for more than a fraction of their daily activities, and when they can observe them, their presence can affect animal behavior. Research on remote monitoring of animal behavior using animal-attached accelerometers has enabled measuring the change in velocity of the body over time and has been used to quantify fine-scale movements and body postures unlimited by visibility, observer bias, or the scale of space use (54). But we clearly cannot use epoxy glue or surgical procedures to attach accelerometers or smartphones to people. Accounting for the various ways how people use their phones, which are consumer grade devices, and propagating the uncertainties involved at different stages of the process, are the most challenging aspects of data analysis. The orientation of the device changes, where people typically carry their phone varies from person to person, some individuals turn their phone off for the night, and so on. None of these complications invalidate inferences drawn from such data, but they do complicate them, which is why the main intellectual challenge in smartphone-based digital phenotyping has arguably moved from data collection to data analysis.

One of the most fundamental problems in the use of smartphones to study human behavior, and one that is nearly always ignored, is whether the phone is on person at different times. For example, the GPS gives the location of the phone, not of the person. These two coincide when the phone is on person, and one approximates the other as long as the phone has recently been observed to be on person. This is similar to the collapse of the wave function in quantum mechanics when the system is observed. Briefly, in quantum mechanics, the so-called wave function, a complex-valued probability amplitude, is used to describe a quantum system. A quantum system is probabilistic, and the wave function can be used to calculate the probabilities of the different measurements that could potentially be obtained from the system when the system is observed. When the system is actually observed, the probability distribution that is associated with observation changes instantaneously and discontinuously; this phenomenon is known as wave function collapse. Translated to our context, consider observing the phone's location from GPS and knowing that the person has the phone on them. One can imagine a probability distribution on a 2D plane, on a map centered at the current location, and there is initially no uncertainty about the location, so the distribution is highly concentrated or peaked. If the person places the phone on the table, as more time passes by, there is greater and greater uncertainty about the location of the person, so the probability distribution, quantifying uncertainty about the location of the person, becomes more and more diffuse. When we next observe the GPS location and if the phone is on the person, this distribution discontinuously jumps to the new location on the map and is again highly peaked, reflecting our current certainty about the location. If the phone

however is not on the person at this time, the probability distribution remains at its original location and continues to grow even more diffuse. The probability distribution keeps evolving in time towards a higher entropy state, unless we observe both the GPS location and know the phone is on the person. Observing when the phone is on person can be done, for example, by using a rule-based classification (e.g., whether there is a call currently in progress), but generally is done most reliably by examining accelerometer data (55).

Missing data is another important consideration in smartphone-based digital phenotyping. Some sensors need to be sampled at different cycles for different reasons, so some missingness is expected by design. For example, GPS can drain the phone battery in a couple of hours, which requires GPS data to be sampled. Furthermore, because the application running on the phone may fail to trigger the GPS for various reasons or the GPS may fail to receive a signal from four satellites needed to ascertain its location, it is critical to be able quantify how successful the sampling process is. The Beiwe platform, for example, alternates between on-cycles and off-cycles of specified lengths when sampling any sensor; if the on-cycle is set to 1 minute and off-cycle to 9 minutes, the sensor is sampled 10% of the time and for each day we would expect to observe 2.4 hours of data. The sampling introduces a challenging missing data problem that must then be addressed. Ignoring the missingness, or simply connecting the last observation of a previous on-cycle with the first observation of the next on-cycle (essentially linear interpolation) can result in a 10-fold error variance for daily summary statistics that are computed from the data and are used in subsequent modeling (56). In the worst case, GPS data may in practice be useless without imputation. A possible alternative to GPS is the use of so-called location service data, which is a method for determining the location of the phone from a combination of various data streams, including cell tower triangulation data and Wi-Fi network data. This method is however proprietary, its precision is unknown, and its resolution varies from location to location, none of which are true for GPS. For this reason, for research purposes, and especially in studies with long follow up, where the accuracy of the imputation improves over time, using GPS data with imputation is the preferred approach.

Transparency and reproducibility apply not only to data collection but also data analysis. Publication of well-documented source code should be the minimum standard. Unlike data, which cannot always be shared to due privacy considerations, there are no arguments from the privacy side against publishing one's code. But sharing code is often not sufficient for successful replication. The first choice in thinking about reproducibility is the choice of language. At the time of writing, Python is arguably the most popular programing language used in data science and related fields. Python itself is open source software that has been copyrighted under a GPL-compatible license certified by the Open Source Initiative (57). This means that analyses implemented in Python can be, at least in theory, run by anyone. In contrast, a language / environment like MATLAB is proprietary and requires a license to use. It is worth pointing out that releasing software under an open source license has implications for intellectual property (IP), and the U.S. legal framework of patent law, copyright, and trade secret is complex and evolving (58).

What complicates the picture is that every Python user takes advantage of several specialized packages, and there are several packages for statistics and machine learning, for example. Because packages typically depend on other packages to function, or on specific versions of specific packages, installing these packages on your own is tedious. This process is made considerably easier by the use of Python distributions that simplify package management and deployment. But distributions do not solve all problems, because the Python code may embed, say, C or C++ code which require their own compilers and linkers. Also, the person who created

the analysis software may have been using a different version / distribution of the operating system, so even after the choice of a programming language, there are many remaining pieces to the puzzle.

Fortunately, we now have very good solutions to these problems. Consider for a moment a traditional computing environment, such as a typical desktop, where we have three layers, from the bottom to the top: hardware, operating system, application. One possibility to improve reproducibility is through *virtualization*, which is a software package that contains a virtual machine and includes an entire operating system, the application, and everything else required to run it. Now in addition to the hardware layer and the operating system layer, we have a so-called hypervisor that creates one or more virtual machines, each having its own operating system and its own applications. As a simplification, we can think of the hypervisor as providing a virtual version of the hardware of the traditional environment, and multiple virtual machines can be run on a single piece of actual physical hardware. Virtualization has its uses, but it requires a lot of resources, which makes it much less appealing as a solution to the reproducibility problem. For replicating analyses, or running a set of analyses on new data, the current best practice is to use *containerization*. Containers are a solution to the problem of how to get software to run reliably from one computing environment (laptop, desktop, cluster, cloud, etc.) to another. A container is a lightweight solution to the problem and consists of the runtime environment that is needed to run the code, meaning the data analysis application or program, all its dependencies, libraries, configuration files, and so on, everything bundled into one package. This makes the container independent of the platform and its dependencies. Currently Docker is the most popular containerization technology. Compared to traditional deployment and virtualized deployment, container deployment has hardware and operating system as its two lower layers, but instead of a hypervisor of a virtualized deployment, it has what is called "container runtime." A container under container deployment is analogous to virtual machine in virtual deployment, but a container does not contain an operating system but instead each shares the operating system kernel with the other containers. This means the containers are lightweight and use fewer resources than virtual machines, making them portable and ideal solutions to the reproducibility problem.

Opportunities in Data Analysis

In order to make smartphone-based digital phenotyping data actionable for interventions, a significant opportunity in data analysis methods is the movement from retrospective analyses carried out after data analysis has been completed to real time or close to real time analyses. This is essential if one is considering implementing interventions in real time based on some changes or anomalies in passive data. These types of approaches might be especially well suited to intervening during high-risk states, such as psychosis, delirium, or suicidality. To use suicide as an example, a recent meta-analysis on the risk factors of suicidal thoughts and behaviors (STB) concludes that little progress has been made towards predicting suicide over a 50-year period (59). The study suggests the need for a shift in focus from risk factors to machine learning-based risk algorithms. Smartphone-based digital phenotyping can at least in principle collect data very close in time to STB, and then what is required are methods that can immediately turn these data into actionable insights.

Computational complexity is an important concept in the quantitative sciences, especially computer science, to quantify how the running time of an algorithm depends on the size of its input. The classic example of computational complexity is that of sorting a list of numbers in ascending order or sorting a collection of

words in lexicographical order (a generalization of alphabetical order). If a person is given 100 numbers and she follows some specific procedure or algorithm, and sorting takes 2 minutes, how long does it take for her to sort 10,000 numbers? The answer depends on the algorithm used and may also depend on how ordered the list is initially. It turns out that there is no sort algorithm that can, for the average case, perform sorting this fast, in so-called linear time (i.e., doubling the input leading to doubling of running time). Some simple sorting algorithms, such as bubble sort, have quadratic computational complexity, meaning that increasing the input 100-fold leads to 10,000-fold computational time. In our example, the person would take approximately 2 weeks to sort the numbers using bubble sort. Using a more efficient algorithm, such as merge sort, which belongs to a difference class of computational complexity, would on average take approximately 22 hours. In most algorithms, there is a tradeoff between performance (computational complexity) and the amount of storage or memory required to carry out the algorithm, which is another important consideration.

Computational complexity is especially important in real time settings. We discussed the importance of imputing missing data for GPS earlier. If the analyses are performed only one time, once data collection has ended, performance is less critical. However, to impute missing data as accurately as possible, say, in settings where we consider each day (a 24-hour period) as our unit of analysis, computational cost begins to compound in the same way as interest compounds on a bank account. To illustrate, to carry out imputation after day 1, the imputation algorithm has to process data from day 1; to carry out imputation after day 2, the imputation algorithm has to process data from days 1 and 2; to carry out imputation after day 3, the imputation algorithm has to process data from days 1 and 2 and 3; and so on; and to carry out imputation after day T, the imputation algorithm has to process data from days 1 through T. Therefore, data from day 1 are processed T times, data from data 2 are processed $T-1$ times, and so on, and data from day T are processed 1 time. This means that repeated computation over time adds an additional linear factor to computational complexity, so that an algorithm that is linear (in the number of days T) when run one time, becomes quadratic when run T times.

This implies that algorithms in these types of settings should generally be chosen based on their computational performance, but one should also aim to develop online versions of these algorithms. Here the term online has nothing to do with the WWW but instead refers to the way the algorithm accommodates previously seen data. In an *offline algorithm*, the algorithm needs the entire dataset to output an answer, whereas an *online algorithm* processes data in a piece-by-piece fashion and generates output along the way. A requirement for doing this is that the algorithm can somehow retain the processed data, or some salient features of the data, in memory in a way that does not depend on the number of data points that have already been processed. An online version of the imputation algorithm would behave in the following way: after day 1, the algorithm processes data from day 1 and generates some summary of the data; after day 2, the algorithm takes the data summary, updates it with data from day 2, and generates the output; after day 3, the algorithm takes the data summary, updates it with data from day 3, and generates the output; and so on; and after day T, the algorithm takes the data summary, updates it with data from day T, and generates the output. In contrast to the offline algorithm, data from each day is processed only once, so the computational complexity does not begin to compound over time. How one can take an offline algorithm and turn it into an online algorithm depends strongly on the algorithm in question and this is in general is very challenging, but it clearly is a tremendous opportunity for digital phenotyping.

Another potential opportunity for digital phenotyping is n-of-1 clinical trials. A clinical trial can be defined as a *prospective study comparing the effects and value of intervention(s) against a control in human beings* (60). To study effectiveness and utility of interventions, in *randomized clinical trials* (RCTs) patients are randomly assigned either to standard care (control group) or to the new intervention (treatment group). The two groups must be sufficiently similar at baseline in relevant respects so that differences in outcome may be attributed to the action of the intervention. Statistical comparisons are made between the two groups, the treatment group and the control group. But because different individuals may respond differently to different interventions, a treatment that works for one person does not necessarily work for another. The practice of investigating group-level response is in stark contrast with clinical practice, where the ultimate end point is the care of individual patients. Single subject or n-of-1 trials have been proposed as the ultimate strategy for individualizing medicine (61), and although they have been used in educational and behavioral settings, they are not commonly used in medicine or public health. Briefly, n-of-1 trials consist of two or more treatments for a single individual where the sequence of the treatments may or may not be randomized. For example, if we denote one treatment with A and another treatment with B, the treatment sequence ABAB constitutes a four-period crossover design (62). Some important considerations in the design of n-of-1 trials include whether one should randomize the sequence, whether the effects of one treatment may be confounded by the carryover effect of a preceding treatment, and whether one should use washout periods between administrations of interventions.

Chronic conditions for which there are easily measurable clinical end points and where the drugs or interventions that are to be tested have a relatively short half-life are most amenable to n-of-1 trials (63). A major challenge in n-of-1 trials is the measurement of relevant end points. Whereas in randomized clinical trials the primary end point might be measured only twice, at the beginning of the trial and at the end of the trial, in n-of-1 trials multiple measurements are needed, and the data collected in these trials is therefore closer to time-series data than traditional longitudinal or panel data. An early paper argued that monitoring and reporting methods in these trials should be as invisible and labor-free to the patient as possible and advocated remote clinical phenotyping and the use of wireless devices for this purpose (64). It was later proposed that cell phones could be used as diaries and one could use actigraphs and other additional devices to learn about outcomes of interest (61). From our point of view, however, the introduction of additional devices implies that the trial is no longer as "invisible and labor-free" to the patient as one would like. In chronic conditions, one would like to be able to follow patients for months or years, and given the generally low adherence of wearable devices, it is unlikely that introduction of new devices, devices that patients themselves would not use or wear if not in a trial, is the most productive route forward. We instead argue for the use of smartphones; in addition to a wristwatch, they are the only technology that we appear to carry on us everywhere. A major opportunity for digital phenotyping is in the development of statistical methods that can inform us about powerful n-of-1 trials, allowing individuals to use different devices (different phones), and yet allow pooling of data so that we can make generalizable population-level claims about the data while at the same time treating each individual patient in the most optimal way.

Future Research Directions

Smartphones appear ideally suited to digital phenotyping given their widespread use and the frequency and extent to which most people use the devices. Wearable devices may one day reach high prevalence, but they

are not there today, and most people stop wearing them after a few months, which is problematic for studying chronic conditions. We have argued that although collection of raw data from smartphones is challenging, and clearly not every study requires that type of data, it is the only approach that makes it possible to aggregate data across studies and investigate aspects of behavior that were not considered at the time the study was run. The downside to the use of raw data is that there are large volumes of data to be moved and the analyses need to be implemented on a (cloud) server. If a study uses Wi-Fi for uploading large data files to economize costs, but the person does not connect to Wi-Fi regularly, the delay may be life-threatening. Therefore, real time applications will necessarily have to use cellular data. In response to the ever greater data consumption of the general public, cell phone companies began deploying the fifth generation technology standard for cellular networks, known as 5G, worldwide in 2019, and it is expected that the new standard will result in substantially faster download and upload speeds, the latter being more critical for smartphone-based digital phenotyping implemented in real time settings.

The main intellectual challenge in digital phenotyping is arguably moving from data collection to data analysis, and we have so far only taken some of the first steps in this area. Just like the development of genome sequencing brought about statistical genetics and genomics, analysis of smartphone data in the biomedical settings will require many methodological developments. Data analysis is difficult enough in experimental settings where every aspect of the process can be tightly controlled, let alone in settings where individuals use various devices in various ways in free living settings. Because of all these degrees of freedom, it is critical that transparency and reproducibility of research are given serious consideration. It is preferable to use highly interpretable models in this endeavor, and ultimately drawing insights from these models requires a strong grounding in human behavior and psychology.

The social, behavioral, and cognitive phenotypes discussed in this review have traditionally been obtained using surveys or clinician-administered tests, which vary greatly among cultures and languages worldwide. Yet humanity is faced with a common set of diseases and our genes are written in a common alphabet. Many have argued that the 21st century may be the century of social and behavioral health, and the ability to measure any phenotype of interest has to be the first step if we wish to manage it. In physics, more precise measurements, such as those as anticipated by Maxwell, led to a scientific revolution that gave rise to quantum mechanics. Although revolutions like that are rare, it is not unrealistic to expect that digital phenotyping could dramatically impact our ability to quantify phenotypes at a global scale, perhaps leading to a quantum leap in our understanding of human disease.

Funding and Disclosure

JPO receives his sole compensation as an employee of Harvard University. His research at the Harvard T.H. Chan School of Public Health is supported by research awards from the National Institutes of Health, Otsuka Pharmaceutical, Boehringer Ingelheim, and Apple. He received an unrestricted gift from Mindstrong Health in 2018. He is a cofounder and board member of a recently established commercial entity that operates in digital phenotyping.

Acknowledgements

I am grateful to my past and present students, postdocs, mentees, mentors, collaborators, and staff for all their hard work, energy, and enthusiasm as we've tackled challenges in the collection and analysis of digital phenotyping data.

Author Contributions

JPO wrote this article.

References

1. Mobile Fact Sheet Pew Research Center2019. Available from: https://www.pewresearch.org/internet/fact-sheet/mobile/.
2. Torous J, Kiang MV, Lorme J, Onnela JP. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. JMIR Ment Health. 2016;3(2):e16. Epub 2016/05/07. doi: 10.2196/mental.5165. PubMed PMID: 27150677; PMCID: PMC4873624.
3. Onnela JP, Rauch SL. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2016;41(7):1691-6. Epub 2016/01/29. doi: 10.1038/npp.2016.7. PubMed PMID: 26818126; PMCID: PMC4869063.
4. Kumar M. Quantum: Einstein, Bohr, and the great debate about the nature of reality: WW Norton & Company; 2008.
5. Delude CM. Deep phenotyping: The details of disease. Nature. 2015;527(7576):S14-5. Epub 2015/11/05. doi: 10.1038/527S14a. PubMed PMID: 26536218.
6. Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási A-L. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences. 2007;104(18):7332-6. doi: 10.1073/pnas.0610245104.
7. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela J-P. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2018;43(8):1660.
8. Lai S, Ruktanonchai NW, Zhou L, Prosper O, Luo W, Floyd JR, Wesolowski A, Santillana M, Zhang C, Du X, Yu H, Tatem AJ. Effect of non-pharmaceutical interventions to contain COVID-19 in China. Nature. 2020. doi: 10.1038/s41586-020-2293-x.
9. Freedman DA. Ecological inference and the ecological fallacy. International Encyclopedia of the social & Behavioral sciences. 1999;6(4027-4030):1-7.
10. Chan Y-FY, Wang P, Rogers L, Tignor N, Zweig M, Hershman SG, Genes N, Scott ER, Krock E, Badgeley M. The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. Nature biotechnology. 2017;35(4):354.
11. Baker J, Barrick E, Eichi HR, Barnett I, Ongur D, Onnela J-P, Buckner R. F242. Intensive Longitudinal Assessment of Mania and Psychosis Using Commonly Available Technologies. Biological Psychiatry. 2018;83(9):S333.
12. Yang WE, Spaulding EM, Lumelsky D, Hung G, Huynh PP, Knowles K, Marvel FA, Vilarino V, Wang J, Shah LM. Strategies for the Successful Implementation of a Novel iPhone Loaner System (iShare) in mHealth Interventions: Prospective Study. JMIR mHealth and uHealth. 2019;7(12):e16391.
13. Reports: Childwise. Available from: http://www.childwise.co.uk/reports.html.
14. media P. Most children own mobile phone by age of seven, study finds The Guardian2020. Available from: https://www.theguardian.com/society/2020/jan/30/most-children-own-mobile-phone-by-age-of-seven-study-finds.
15. Kiang MV, Chen J, Krieger N, O Buckee C, Onnela J-P. Working Paper: Human Factors and Missing Data in Digital Phenotyping. 2020.

16. Lupač P. Digital Divide Research', Beyond the Digital Divide: Contextualizing the Information Society: Emerald Publishing Limited; 2018.
17. Noah B, Keller MS, Mosadeghi S, Stein L, Johl S, Delshad S, Tashjian VC, Lew D, Kwan JT, Jusufagic A. Impact of remote patient monitoring on clinical outcomes: an updated meta-analysis of randomized controlled trials. NPJ digital medicine. 2018;1(1):1-12.
18. Dorsch AK, Thomas S, Xu X, Kaiser W, Dobkin BH. SIRRACT: an international randomized clinical trial of activity feedback during inpatient stroke rehabilitation enabled by wireless sensing. Neurorehabilitation and neural repair. 2015;29(5):407-15.
19. Logan AG, Irvine MJ, McIsaac WJ, Tisler A, Rossos PG, Easty A, Feig DS, Cafazzo JA. Effect of home blood pressure telemonitoring with self-care support on uncontrolled systolic hypertension in diabetics. Hypertension. 2012;60(1):51-7.
20. Klasnja P, Consolvo S, Pratt W, editors. How to evaluate technologies for health behavior change in HCI research. Proceedings of the SIGCHI conference on human factors in computing systems; 2011.
21. Shih PC, Han K, Poole ES, Rosson MB, Carroll JM. Use and adoption challenges of wearable activity trackers. IConference 2015 Proceedings. 2015.
22. Partners E. Inside Wearables: How the Science of Human Behavior Change Offers the Secret to Long-term Engagement. 2014.
23. Gartner. User Survey Analysis: Wearables Need to Be More Useful2016.
24. Khan WZ, Xiang Y, Aalsalem MY, Arshad Q. Mobile phone sensing systems: A survey. IEEE Communications Surveys & Tutorials. 2012;15(1):402-27.
25. EmotionSense 2020. Available from: http://www.emotionsense.org/
26. Miluzzo E, Cornelius CT, Ramaswamy A, Choudhury T, Liu Z, Campbell AT, editors. Darwin phones: the evolution of sensing and inference on mobile phones. Proceedings of the 8th international conference on Mobile systems, applications, and services; 2010.
27. Mobile phone based sensing software Wikipedia. Available from: https://en.wikipedia.org/wiki/Mobile_phone_based_sensing_software.
28. The average 'good' indie game makes just $25,000 in its first year on sale, says Grey Alien's Birkett 2018. Available from: https://www.pcgamesinsider.biz/interviews-and-opinion/66655/the-average-good-indie-game-makes-just-25000-in-its-first-year-on-sale-says-grey-aliens-birkett/.
29. Overview. Available from: http://researchkit.org/docs/docs/Overview/GuideOverview.html.
30. Timed Walk. Available from: http://researchkit.org/docs/docs/ActiveTasks/ActiveTasks.html#timed.
31. CMSensorRecorder Apple Developer. Available from: https://developer.apple.com/documentation/coremotion/cmsensorrecorder.
32. iPhone Users Spend $101 Every Month on Tech Purchases, Nearly Double of Android Users, According to a Survey Conducted by Slickdeals PR Newswire2018. Available from: https://www.prnewswire.com/news-releases/iphone-users-spend-101-every-month-on-tech-purchases-nearly-double-of-android-users-according-to-a-survey-conducted-by-slickdeals-300739582.html?c=n
33. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011;10(9):712-.
34. Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012;483(7391):531-3.
35. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015;13(6):e1002165.
36. Koo BM, Vizer LM. Mobile technology for cognitive assessment of older adults: a scoping review. Innovation in aging. 2019;3(1):igy038.
37. Dhand A, Lang CE, Luke DA, Kim A, Li K, McCafferty L, Mu Y, Rosner B, Feske SK, Lee J-M. Social Network Mapping and Functional Recovery Within 6 Months of Ischemic Stroke. Neurorehabilitation and neural repair. 2019;33(11):922-32.
38. Hudson K, Lifton R, Patrick-Lake B. The precision medicine initiative cohort program—Building a Research Foundation for 21st Century Medicine. 2015.

39. Notice of Information: NIMH High-Priority Areas for Research on Digital Health Technology to Advance Assessment, Detection, Prevention, Treatment, and Delivery of Services for Mental Health Conditions. National Institute of Mental Health; 2018.
40. National Advisory Mental Health Council (NAMHC) [cited 2020]. Available from: https://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/index.shtml.
41. Health TNIoM. Research Domain Criteria (RDoC). Available from: https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/index.shtml.
42. Reis HT. Why researchers should think" real-world": A conceptual rationale2012.
43. Koster EH, De Raedt R, Leyman L, De Lissnyder E. Mood-congruent attention and memory bias in dysphoria: Exploring the coherence among information-processing biases. Behaviour research and therapy. 2010;48(3):219-25.
44. Myin-Germeys I, Klippel A, Steinhart H, Reininghaus U. Ecological momentary interventions in psychiatry. Current opinion in psychiatry. 2016;29(4):258-63.
45. Solhan MB, Trull TJ, Jahng S, Wood PK. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychological assessment. 2009;21(3):425.
46. Bos FM, Schoevers RA, aan het Rot M. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. European Neuropsychopharmacology. 2015;25(11):1853-64.
47. aan het Rot M, Hogenelst K, Schoevers RA. Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. Clinical psychology review. 2012;32(6):510-23.
48. Oorschot M, Kwapil T, Delespaul P, Myin-Germeys I. Momentary assessment research in psychosis. Psychological assessment. 2009;21(4):498.
49. Shiffman S. Ecological momentary assessment (EMA) in studies of substance use. Psychological assessment. 2009;21(4):486.
50. Walz LC, Nauta MH, aan het Rot M. Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: A systematic review. Journal of anxiety disorders. 2014;28(8):925-37.
51. Haedt-Matt AA, Keel PK. Revisiting the affect regulation model of binge eating: a meta-analysis of studies using ecological momentary assessment. Psychological bulletin. 2011;137(4):660.
52. Linux Servers Under Attack for a Decade 2020. Available from: https://www.infosecurity-magazine.com/news/linux-servers-under-attack-for-a/.
53. Schryen G. Security of open source and closed source software: An empirical comparison of published vulnerabilities. AMCIS 2009 Proceedings. 2009:387.
54. Brown DD, Kays R, Wikelski M, Wilson R, Klimley AP. Observing the unwatchable through acceleration logging of animal behavior. Animal Biotelemetry. 2013;1(1):20.
55. Barback J, Onnela JP. Working paper: Accelerometry-based algorithm for smartphone proximity classification. 2020.
56. Barnett I, Onnela J-P. Inferring mobility measures from GPS traces with missing data. Biostatistics. 2020;21(2):e98-e112.
57. Python Copyright. Available from: https://www.python.org/doc/copyright/.
58. Jin HR. Think Big: The Need for Patent Rights in the Era of Big Data and Machine Learning. NYU J Intell Prop & Ent L. 2017;7:78.
59. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, Musacchio KM, Jaroszewski AC, Chang BP, Nock MK. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychological bulletin. 2017;143(2):187.
60. Friedman LM, Furberg C, DeMets DL, Reboussin DM, Granger CB. Fundamentals of clinical trials: Springer; 2010.
61. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Personalized medicine. 2011;8(2):161-73.

62. Barlow DH, Nock M, Hersen M. Single case experimental designs: Strategies for studying behavior for change2009.
63. Guyatt GH, Heyting A, Jaeschke R, Keller J, Adachi JD, Roberts RS. N of 1 randomized trials for investigating new drugs. Controlled clinical trials. 1990;11(2):88-100.
64. Topol EJ. Transforming medicine via digital innovation. Science translational medicine. 2010;2(16):16cm4-cm4.

Figure and Table Legends

None

Tables

None

Figures

None

Boxes

None