PERSPECTIVE

Data Science Approaches for Effective Use of Mobile Device–Based Collection of Real-World Data

Larsson Omberg[1,*], Elias Chaibub Neto[1,*] and Lara M. Mangravite[1,*]

The use of mobile health for monitoring disease outside of the clinic has opened new opportunities for drug testing and monitoring. In particular, these tools are providing new experimental designs for collection of real-world data. These technologies and queries, although promising, require the application of analytical methods that can accommodate the uncontrolled, unmonitored, individualized, and, often, near continuous data streams. Here, we discuss opportunities and ramifications on analytical considerations.

Mobile health, that is, the evaluation of health outside of the clinic using wearables and smartphones, and, more broadly, the collection of real-world evidence,[1] provide opportunities to advance multiple goals for monitoring drug response, including the monitoring of efficacy through digital biomarkers that can be used as primary end points for drug efficacy, monitoring of patient-reported outcomes and/or quality of life measures, and of toxicities and/or response to long-term exposures. Although digital end points are of interest to regulatory agencies such as the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) and are starting to be integrated as primary end points into clinical trials,[2] adoption is slow. In part, this is due to difficulties in quantifying the accuracy of measures when they are collected in an unmonitored manner and in an uncontrolled setting. Indeed, the ability to develop robust measures that are reliably accurate requires both an expanded validation plan designed to pressure-test the measure across a range of conditions and a good understanding of the impact that variations in daily living can have on data collection. Because interpretation of mobile health data involves the processing and analysis of high dimensional, longitudinal sensor data collected in continuous or near continuous data streams, it requires the use of statistical approaches that account for repeat measures as well as extensive use of signal processing and/or machine-learning techniques. These approaches provide opportunity for sensitive, individualized monitoring of drug responses. Here, we provide a short introduction to the importance of appropriate usage of analytic and machine-learning techniques for the interpretation of mobile health data (see also refs. 3–5). This includes a description of the types of experiments and data that can be collected using mobile health and some examples from the literature that highlight important analytical considerations. Although these observations are relevant to any device that is collecting sensor data in a continuous or near continuous manner, we exemplify these issues using our own experience with the development and analysis of smartphone-based measures.

FROM SENSORS TO MACHINE LEARNING AND SOME CONSEQUENCES

Modern smartphones have opened a wealth of possibilities to extend electronic health monitoring for two reasons: (i) The always connected nature and computational power of smartphones allows for rapid data collection and (ii) the large number of embedded sensors allows for multimodal data collection.[6] A typical phone has sensors that can measure acceleration, rotation rates, magnetic fields, sound levels, record audio and video, and record time and touch through the screen, among other capabilities. Sensor-based data collection performed in the context of protocols, or tasks designed to capture disease relevant behavior, can be used to generate hundreds of phenotypic measurements, including those that mimic evaluations typically performed in the clinic (e.g., sit to stand test for mobility or blood pressure measurement). They can also be used to passively collect measurements during daily activities (e.g., mobility analysis performed while an individual is walking). In

[1]Sage Bionetworks, Seattle, Washington, USA. *Correspondence: Larsson Omberg (larsson.omberg@sagebionetworks.org), Elias Chaibub Neto (elias.chaibub.neto@sagebionetworks.org), and Lara M. Mangravite (lara.mangravite@sagebionetworks.org)

Received November 18, 2019; accepted January 13, 2020. doi:10.1002/cpt.1781

either case, high dimensional data streams are generated that require extensive processing and analysis to be converted into phenotypic measures. For gait analysis, inertial measuring units embedded in wearables and phones collect time series data consisting of 100 Hz recordings on 6 axes (three from the accelerometer and three from the gyroscope). Data such as these can be analyzed in three ways. First, features with established clinical relevance (e.g., gait speed) can be extracted through signal processing. For this approach, algorithms are manually evaluated and tuned to maximally approximate the desired phenotypic measure. Although this first approach provides measures with clear clinical interpretation, it can limit use of the full spectrum of information provided in the collected data. To address this, one could opt to use traditional machine learning in conjunction with signal processing to select a subset of promising features from a larger set of exploratory features generated by signal processing. This data-driven approach might be better able to distinguish disease state across heterogeneous populations, as it works by optimizing on the outcome of interest. In the case of gait, this second approach is suited to identifying a broader set of gait disturbances in addition to gait speed. Finally, machine-learning methods based on deep-learning models have also been used to generate features in an automatic and data-driven way, bypassing the need for signal processing.

Although machine-learning approaches provide the opportunity to develop more comprehensive digital measures, the use of machine learning must be done appropriately in order to avoid subtle errors. Because they are data driven, machine-learning approaches will leverage any source of variation in a dataset, including variability due to biology, technical artifacts, and even random noise (especially in small datasets). Identification of biologically relevant measures requires disciplined analysis. This is typically addressed by using two datasets—training data is used to train models and select potential features, whereas a separate validation dataset, assumed to contain similar biological but different technical variation, is used to evaluate the predictive performance of the trained model and confirm the biological relevance of the new features (e.g., by comparing then to existing clinically validated outcomes or severity measures). Because two datasets are not always available in mobile health studies, a single dataset is often split to support both training and validation functions. This can be problematic for small datasets. There are many papers reporting positive validation results of digital measures. Many of these results are developed using machine learning in small sample size studies, which can promote exaggerated results that will not replicate in other datasets. This is best addressed by reporting the uncertainty in measure performance. As an example, a study reporting diagnostic accuracy using the area under the receiver operating curve should be expected to report error bars as a means to help readers understand the uncertainty in the reported performance. In addition, the choice of performance metric is dependent on the nature of the data. Reporting on the incorrect metric (e.g., area under the receiver operating curve in extremely imbalanced datasets) can lead to inflated interpretation of accuracy.[7]

APPLICATION AND CONSEQUENCES OF LONGITUDINAL SAMPLING

A major benefit of mobile health is the opportunity to tailor health monitoring to each individual. This is of particular benefit for conditions and treatments that present in a highly heterogeneous manner across individuals or change dynamically over time. Because mobile health provides longitudinal data collection with frequent sampling, it can be used to capture individualized changes over time by using personalized models or n-of-1 analysis.[3] Analysis of frequently sampled longitudinal data requires an analytical approach that is distinct from those used for sparsely sampled data. Although repeated measures collected from an individual are autocorrelated, a common mistake observed in the literature is to assume that these repeated measures are independent. If not taken into account, autocorrelation can lead to an inaccurate estimate of the number of false-positive discoveries in an analysis. Notably, this can result in either an underestimate or an overestimate depending on whether the autocorrelation is positive or negative.[3,4] Furthermore, the incorrect use of the repeated measurements in population level analysis, such as classification studies can lead to *identity confounding* artifacts, where the classifier is mostly distinguishing differences across individuals instead of differences across conditions or disease states. A recent literature review of mobile health classification studies demonstrated that 47% had artificially inflated the performance of their measures through failure to account for the identity of individual data points.[8] Our own quantification of this effect across three studies showed that identity confounding can be many times larger than the effect of the condition that was being studied.[9] As with the analytical issues described above, proper interpretation of analyses using mobile health studies for classification requires reporting of how repeat measures were handled.

POSTMARKET MONITORING, OPEN ENROLLMENT, AND THE EFFECTS OF CONFOUNDERS

Fully remote mobile health studies can support low cost enrollment of large swaths of the population as compared to in-clinic studies. Many studies relying solely on mobile health measures have enrolled in the tens of thousands from across distributed geographic regions, providing the opportunity for broad sampling across diverse populations in the real-world setting. This approach can be a good option for postmarket monitoring studies, including to evaluate real-world drug efficacy and toxicity as well as market fit. It can also be used to prescreen for enrollees into clinical trials. In these contexts, data are often collected using open enrollment techniques. Because these can lead to biased sampling of the population, they must be carefully evaluated and interpreted. For example, we recently recruited 17,000 individuals into a Parkinson's disease study using an open enrollment approach. The control population tended to be significantly younger than the Parkinson's disease population (average age 38 vs. 61). Because age was correlated with disease status, machine-learning methods could trivially distinguish between cases and controls by selecting features related to age rather than those related to disease state. With careful consideration these issues can be both assessed and accounted for.[10]

In this case, we did so by rebalancing the populations according to clinical covariates and by measuring performance of the classifier both before and after correction for known covariates.[10]

CONCLUSIONS

The use of mobile health to collect frequent measures in a real-world setting provides a promising tool to aid in drug development and monitoring. Appropriate use and interpretation of these approaches, which also provide great opportunity to monitor lived experience, require careful attention to analytical techniques. Much of the success of mHealth will be dependent on comprehensive validation of developed measure and objective benchmarking of analytical techniques used in their interpretation. With appropriate application, these approaches stand to greatly advance our ability to objectively assess the impact of treatments on individuals' lives.

FUNDING

No funding was received for this work.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

1. Sherman, R.E. *et al*. Real-world evidence - what is it and what can it tell us? *N. Engl. J. Med.* **375**, 2293–2297 (2016).
2. Goldsack, J. *et al*. Digital endpoints library can aid clinical trials for new medicines - STAT. STAT <https://www.statnews.com/2019/11/06/digital-endpoints-library-clinical-trials-drug-development/> (2019).
3. Chaibub Neto, E. *et al*. On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: the mPower case study <https://ui.adsabs.harvard.edu/abs/2017arXiv170609574C/abstract> (2017).
4. Barnett, I., Torous, J., Staples, P., Keshavan, M. & Onnela, J.-P. Beyond smartphones and sensors: choosing appropriate statistical methods for the analysis of longitudinal data. *J. Am. Med. Inform. Assoc.* **25**, 1669–1674 (2018).
5. Hicks, J.L. *et al*. Best practices for analyzing large-scale health data from wearables and smartphone apps. *NPJ Digit. Med.* **2**, 45 (2019).
6. Perry, B. *et al*. Use of mobile devices to measure outcomes in clinical research, 2010–2016: a systematic literature review. *Digital Biomarkers* **2**, 11–30 (2018).
7. Cokelaer, T. *et al*. DREAMTools: a Python package for scoring collaborative challenges. *F1000Res*. **4**, 1030 (2015).
8. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C. & Kording, K.P. The need to approximate the use-case in clinical machine learning. *GigaScience* **6**, 1–9 (2017).
9. Chaibub Neto, E. *et al*. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ Digit. Med.* **2**, 99 (2019).
10. Chaibub Neto, E., Tummalacherla, M., Mangravite, L. & Omberg, L.Causality-based tests to detect the influence of confounders on mobile health diagnostic applications: a comparison with restricted permutations <https://arxiv.org/abs/1911.05139> (2019).