Case Example: Optimizing Data Quality and Participant Privacy

We designed this exploratory observational study to collect a variety of digital data from approximately 150 participants with a diagnosed respiratory condition in order to develop and test predictive models for symptom exacerbation. The ultimate goal was to design a digital/mHealth framework that will allow patients to better understand and manage their disease.

We planned to collect diverse data including activity, heart rate, environmental factors, rescue inhaler usage, respiration rate, and patient reported outcomes (PRO) from patients using sensors digital technologies.

Collecting Data While Protecting Privacy

Recording participants' environmental factors—including temperature, humidity, and a variety of air quality metrics—was critical to the success of this study. Participants' GPS coordinates were required for obtaining environmental factors. However, as recording GPS coordinates could compromise participant privacy, the cell phone application was designed to only store the environmental factor data, with no permanent record of the GPS coordinates.

Demonstrating Data Parsimony

To avoid data overload, 1) we collected only data for variables relevant to the clinical question, and 2) selected the optimal level of detail for data collection. Findings from previous research, along with input from clinicians and patients, informed the variables included for data collection from the sensors and digital technologies.

To determine optimal data granularity, different smoothing algorithms were tested to inform what ***epoch length*** best eliminated noise without obscuring clinically relevant signals. Applying smoothing algorithms with optimal ***epoch length*** had the following benefits: 1) minimized the amount of streaming data to be stored and transmitted, 2) maximized the signal to noise ratio, and 3) data were better interpretable with appropriate time units.

Minimizing Missing Data

Prior to writing this protocol, we ran a pilot study to compare the available technologies and to select the optimal ones. In this phase, healthy volunteers were asked to use 2 technologies each for a period of 4 weeks. Participant adherence was extremely poor. In response, the study protocol was written in a manner to promote adherence and reduce missing data.

First, in response to feedback from pilot participants, the decision was made to share participants' data with them during the study in order to promote engagement and adherence. There were no concerns that this decision would threaten the study objective: to evaluate whether we can capture reasonably good quality and complete data, and if the data could be used for model development. To ensure that participants could understand and find value in the data shared with them, patient surveys and beta testing were used to develop an intuitive, user-friendly, visual interface to provide summary feedback to participants.

Second, automated, centralized monitoring was planned with two layers of action in response to missing data. If a participant did not generate PRO data for a certain number of days (say "n"), they would be prompted to do so by an automated message generated by the application custom developed for this study. We also designed a portal system to continuously monitor data

capture from all connected technologies. We planned that participants should be contacted by a third party to understand the reason for any disengagement if the data is not generated in the portal for a sustained period of time ("m" number of days, where m > n). The contact was intended to either 1) re-engage the participant, or 2) address any potential issue that the subject was encountering, such as technology malfunction.

Finally, while designing the study, a nominal financial incentive was tied to a subject's adherence rate. In determining the exact dollar amount of the incentive, factors considered included 1) fair market rate, 2) the risk of the cash incentive becoming a perverse incentive, affecting data attribution, and 3) the amount that would provide sufficient incentive for adherence. Patients were engaged in this decision.

Positive Results

This multipronged approach was designed to yield significantly higher adherence rates during the study.

Reference: Relevant CTTI Considerations

For additional considerations pertaining to data quality, please reference CTTI Recommendations for Managing Data.

- Sponsors should ensure that appropriate ***meta-data*** is collected to provide sufficient contextual information required to understand the outcome data captured by digital technologies and allow it to be readily interpreted while avoiding the collection of intrusive data. (Click here for more)
- The principle of data parsimony is particularly important. (Click here for more)
- When using digital technologies for data capture, a multi-pronged approach to preventing missing data is optimal, with efforts focused on 1) optimizing trial design, 2) ensuring technical approaches are in place to eliminate any technology-or transmission-related causes of missing data, and 3) pilot testing to identify any unanticipated causes of missing data. (Click here for more)