

## 5. Protocol for Statistical Analysis

Statistical analysis for economic evaluation of pediatric telehealth interventions differs from statistical analysis used for studies that measure a clinical response to a specific, controlled intervention (e.g., a drug trial). The foundation of good statistical analysis is a good study design, with detailed understanding of the intervention group, the comparison group (which could be the same patients in the pre-intervention time period, a matched cohort that did not receive the intervention, or potentially both), and time period for which data will be collected. Other items that should be determined before starting the study:

- Sample size/power analysis
- Data sources available and their inherent limitations (see document #4 in this framework on Design Issues and Data Sources)
- Clear operational definitions for each variable in the dataset
- Covariate data that are available in the chosen data sources. If using different data sources, understand that certain data (especially race and ethnicity) unfortunately may not be identical for the same patient in each data source. For example, one data source may allow multiple ethnicities to be entered for a single patient whereas another data source may only allow a single ethnicity to be entered for each patient. If you pull ethnicity data from one database and use it to do multivariable analysis on data from several databases, you risk statistical error unless you are certain that any differences are minimal.

Due to the complexity of statistical analysis protocols for these studies, we provide an example here of a design and analysis narrative. This study was a comprehensive economic analysis of a program that used a multi-component telehealth intervention to reduce health care utilization for medically complex children. This level of detail may not be necessary for your study, but it provides a guide for items to consider:

Example of a design and analysis narrative (tailor wording to your study):

### **Study Design**

*This retrospective observational study will use a pre-post design with a contemporaneous control group to compare utilization and costs of care, with each treated patients to three controls.*

### **Population and Data Extraction**

*We will provide a “finder file” with demographic information (patient name birthdate sex, race and address) on our patient group to the public agency (Agency) responsible for collection of the required UB04 (billing) data for ED and hospital admissions in our state. The Agency staff will link our records to their archival billing data, place a binary variable (STUDY\_PATIENT=1) and remove all identifying variables from this file to meet the criteria for Non-human Research. The Agency staff will then select a random sample of patients by age, sex and race distribution. Five control patients will be selected for each study patient (extras will assure a good final 1:3 match). Agency staff will add the binary variable STUDY\_PATIENT=0 to these control group data. The Agency will transfer these records through encrypted FTP or other approved method to the Principal Investigator (PI) and research analyst (RA).*

### **Dataset Construction**

*The PI and RA will receive and examine the deidentified data for data quality and missingness. Data will be separated in four data sets for constructing the analytical files: 1) pre-period patient, 2) post-period patient, 3) pre-period control, and 4) post-period control. Baseline demographic information will be constructed for each subject as well as indicators of disease severity and comorbidity burden. Summary values of utilization and cost (or charges) will be constructed for each time period. For the analytical data set generation, study patients will be propensity-score matched 1:3 to control patients using baseline characteristics. The matching will be performed using Proc PSMATCH in SAS Version 9.4. We will use a Greedy Matched Algorithm with a caliper setting of 0.2.*

### **Study Power and Statistical Analysis**

#### **Sample Size and Power:**

*Sample size is prescribed by the patients enrolled in the program during the specified time period, we estimate up to 100 study patients will be included in the study with 300 controls. This small number may make it difficult to find statistically*

## 5. Protocol for Statistical Analysis

*significant differences between the pre- and post-intervention period for measures with large variances. This is especially the case for documenting cost savings, because costs vary greatly and thus require large samples.*

*The primary study outcome is differences in ED visits and hospital overnights for the pre-and post-enrollment periods. Assuming 74 total days per annum with an ED visit, outpatient surgery or inpatient hospital overnight in the pre-intervention year and a 10% reduction in in the post-intervention time we will have > 99% power to detect an improvement. This estimate assumes a 2-sided comparison of two event rates in terms of person years with a sample of 90 person-years per group at the 0.05 significance and a difference in event rates of at least 7 if the control group based on a pre-event rate of 74.*

*For cost within the population, with a sample size of up to 100 patients, it will be essential to control for cost drivers that increase the variation in costs within the population. With this sample size, we will have >80% power to detect a 33% decrease in cost of care from the pre to post period. This estimate is based a two-sided test of the comparison of two gamma means. The power estimate assumes we will be using a gamma regression generalized linear model. Under this assumption, a pre-enrollment sample size of 100 and a post-enrollment sample size of 100 cases achieves 80.84% power at the 5% significance level when pre and post enrollment means are \$200,000 and \$134,000 respectively. This is based on the pre-enrollment shape parameter being 1 and the post-enrollment shape parameter being 1. However, using statistical controls of demographic and disease severity measures may be expected to improve our power enough for us to detect a statistical difference for cost savings of about 15-20%.*

### **Statistical Methods:**

*Means, medians, and standard deviations will be used to estimate unadjusted costs and days at home. Frequencies and percentages will be calculated to describe hospitalizations and ED visits in the pre and post time periods. To test for differences in the proportion of patients admitted by risk groups defined by a summary score of disease severity and comorbidity burden variables constructed from the UB04 ICD10 codes, a chi-square statistic will be calculated. Testing for differences between groups regarding cost and days at home, a t-test or Wilcoxon signed-rank will be calculated as appropriate. Multivariable regression methods will be used to control for number of patient days in the follow-up period and aggregate measures of risk factors that are potentially related to the outcome. Covariates selection will be adjusted as needed but may include age, gender, summary measure of disease severity and summary comorbidity score. Covariates not found to significantly contribute to each model will be removed, manually, one at a time and the models re-fit to the data. Clinically relevant variables chosen based on expert team opinion and peer-reviewed evidence will be used to determine which covariates will be initially included in the models to control for potential confounding due to emergence of new conditions in the post period.*

*Generalized linear modeling will be used to test the hypothesis that the ED, hospital days and costs of healthcare are different for the pre and post time period. Poisson distributions will be used for modeling count data. To correct for the non-normal distribution of healthcare costs, gamma distributed generalized linear models using a logarithmic transformation will be used after confirming good model fit. The use of a gamma distributed generalized linear model with a log transformed link function has been shown to be a good method to estimate healthcare cost distributions that are generally right-skewed, especially when the log transformed dependent variables do not have heavy tails or excessive heteroscedasticity.*

*The use of archival data required for reimbursement decreases the risk of informative missingness. Where data are missing, we will examine the complete case analysis and compare to an imputed analysis set using multiple imputation with the assumption of missing at random.*