Data Processing, Cleaning, and Creation Outline

1. Data from relational database (eventually we will connect directly)
   1. All registry data: vregallYEARMMDD.Rda
      1. All potential recruits for some sites, but not for all sites
   2. All survey data: vsurvallYEARMMDD.Rda
   3. Merged registry and survey data: vcomballYEARMMDD.Rda - eventually this will be the main data set for all processing
   4. Johns Hopkins data- subset of individuals combined (registry/survey) merged with Tamara’s IDs: not sure we will need this in the future
   5. Location: DataProcesssingCleaningCreation/Data
2. Survey
   1. Survey Processing (Survey\_Processing\_2022\_03\_02.Rmd)
      1. Initial data processing to assess values etc.
      2. Provides report output but does not alter data
      3. Location: DataProcessingCleaningCreation/Survey/
      4. Status: done (03/29/22)
   2. Survey Cleaning(Survey\_Cleaning\_2022\_03\_02.Rmd)
      1. Cleaning of variables, variables
      2. Provides report and creates a vsurvall\_cleaned\_YEARMMDD.Rda
      3. Location: DataProcessingCleaningCreation/Survey/
      4. Status:
         1. adding cleaning functions (03/29/22)
   3. Survey Create Variables (Survey\_CreateVariables\_2022\_03\_02.Rmd)
      1. Create composite variables that use original survey variables
      2. Provides report and creates new variables and file vsurvall\_analysis\_YEARMMDD.Rda
      3. Location: DataProcessingCleaningCreation/Survey/
      4. Status
         1. Update create variables: see original clean and create report (03/29/22)
3. Registry
   1. Registry processing
      1. Initial data processing to assess values, etc.
      2. Provides report output but does not alter data
      3. Location: DataProcessingCleaningCreation/Registry/
      4. Status: Done (03/29/22)
   2. Registry cleaning and creation
      1. Cleans and creates composite variables that use original registry variables
      2. Provides report and creates new variables and file vregall\_analysis\_YEARMMDD.Rda
      3. Location: DataProcessingCleaningCreation/Registry/
      4. Status (03/29/22)
         1. Continue to clean and add new variables (e.g additional risk scores; treatment variables)
         2. Add functionality of table summaries by site
4. Merged Data: to be done
5. R-shiny:

a. Code to take Analysis data from merged and display… we have template but need to develop further