**Task Based Sampling Code**

**Natalie Baughan, MIDRC TDP 3d**

**April 2023**

**Updated January 2024, Heather Whitney**

This folder contains all associated scripts and spreadsheets for running and evaluating the developed task based sampling process.

The base folder contains the standardized functions and blank spreadsheets that will be used in each implementation. Copy and rename blank spreadsheets for each unique use.

The subfolder Example\_TaskBasedSampling contains all files for walking through an implementation of task based sampling using data from the public MIDRC data commons, and the process for this example is outlined below.

**Task Based Sample Example**

1. Log into data exploration tool at <https://data.midrc.org/explorer>
2. Navigate to imaging studies and select your cohort
   1. Example: Under LOINC Properties, select Method (Modality) = ‘CT’ and System (Body Region) = ‘Chest’
   2. Save to chosen input data location with specified name
   3. The tutorial uses the data available at <https://doi.org/10.60701/P67C-YW55>.
3. Specify your target population
   1. Begin with the ‘MIDRC\_TaskBasedSample\_TargetSpecification.xlsx’ spreadsheet and specify your target population demographic distributions in decimal percentages (0.5 = 50%) for the categories of age, race, sex at birth, ethnicity, and COVID-19 status.
      1. This example uses an approximate CDC case distribution with a 50:50 COVID-19 status as shown in “MIDRC\_TaskBasedSample\_TargetSpecification\_DemoExample.xlsx”
   2. NOTE: The ‘MIDRC\_TaskBasedSample\_TargetSpecification.xlsx’ spreadsheet contains 26 default subcategories for the four demographic categories listed above. These categories are defined in alignment with the data model at data.MIDRC.org. If you would like to modify these categories, do be aware that the associated scripts will need to be modified correspondingly, and the data categories provided from metadata files at data.MIDRC.org may not align with newly defined categories.

*Within your script:*

*(Sections in the code should correspond to sections described below)*

1. **Read in Data:** Modify the auxiliary script ‘Call\_TaskBasedSample\_Example.m’ to read in your metadata input file and modified target specification spreadsheet.
   1. NOTE: If you get a “Unknown variable name: 'age\_at\_index' error”
      1. Key variables ‘age\_at\_index’, ‘race’, ‘sex\_at\_birth’, ‘ethnicity’, ‘covid19\_positive’, and ‘study\_modality’ may be labeled in the downloaded metadata sheet with a trailing ‘\_0’ after the variable name. You will need to remove the trailing ‘\_0’ and re-save the spreadsheet to ensure the script is able to read in the variable correctly.
   2. The script will present a warning about duplicate patients within the metadata file. It is best to remove these after filtering major items and before initiating the optimization. Leaving duplicate patients in the optimization will lead to misleading results since demographic characteristic will be represented twice.
2. **Clean Data #: …”** The following code sections are for data cleaning, to ensure all key variables are labeled uniformly.
3. **Filter Major Items:** This is the primary section to further modify your dataset for anything not selected or selectable on data.MIDRC.org. Examples show how to perform this in a variety of way using [logical indexing](https://www.mathworks.com/help/matlab/matlab_prog/access-data-in-a-table.html#mw_77d04ebd-b6f7-4ee9-93ac-1cc1fd7bc735) on selected variables.
4. **Take only unique patients:** This section will remove any remaining duplicate patients from your input table.
5. **Specify initial distributions:** This section calls the ‘CountTBSCategories.m’ function to quantify the initial demographic distribution of your data.
6. **Specify target:** The section reads in your target specification spreadsheet to be used in optimization.
7. **Call task based sample:** This section calls the ‘TaskBasedSample.m’ function to perform the optimized quota sampling and select a subset of your input data matched to your specified target.
   1. **Input variables:**
      1. inputTable: Generally, leave as current variable ‘data’
      2. inputTarget: Matrix defined by target specification spreadsheet
      3. threshold: Maximum percent deviation allowable in any demographic category between current sample and target specification. Typical values = 0.1 ~ 0.05 (5% - 10% maximum deviation). This variable is used to track the optimization process and stop the optimization when a “sufficient” sample is found. Plotted in Figure 2 if mkfigures = 1.
      4. mkFigures: Set = 1 if you would like to generate explanation figures or = 0 if not. Three figures should appear following optimization.
         1. Figure 1 – Histogram of patient demographic fit scores. In each iteration, all patients in the sample are assigned a score based upon the joint weighting of their demographic profile. Histogram of all patient scores is generated at iteration number 10, 50 (if reached), and end. Histogram should get narrower with increasing iteration but will never be all at 0.

Chart

Description automatically generated

* + - 1. Figure 2 - Maximum percent deviation as described above, calculated at each iteration. This metric can be interpreted similar to a loss function of the optimization. Should decrease over time until threshold is met. Discontinuities in the plot slope indicate a shift from one category dominating the max percent deviation to another category, as subjects are removed from the sample and overall demographic distributions shift.

Chart

Description automatically generated

* + - * 1. If this metric plots erratically or unexpectedly increases, the demographic distribution of the sample is shifting too fast for the optimization. You will need to decrease the patients removed in each iteration or begin the optimization with more patients.
      1. Figure 3 – Sample population. Should continually decrease until optimization is complete.

Chart, line chart

Description automatically generated

1. **Specify final distributions:** This section calls the ‘CountTBSCategories.m’ function to quantify the final demographic distribution of your data and concatenates the initial and final distributions for post-optimization review.
2. **Write results sheet:** This section writes out the target specification, initial and final demographic distributions of your data to a template Excel sheet for review. Prior to running this section, please copy the ‘MIDRC\_TaskBasedSampleResults\_Blank.xlsx’ sheet and rename it to your sample, the copy you specify in the code will be populated with your optimization results.
3. **Write results output file:** This section writes a spreadsheet of the patients selected to be in your final sample. All initial columns will be included. This file can be used to filter patients downloaded from the data explorer using methods of the user’s choice.

