STS extraction

September 6, 2024

A Docker container to anonymize/recode STS (Society of Thoracic Surgeons) data

Nicholas Ollberding, Ph.D. Division of Biomedical Informatics Cincinnati Children's Hospital Medical Center
November, 2024

Introduction

The STS docker container is used to strip PHI from STS data file and remove procedures occurring in unconsented participants > 18 years of age

After installation, the software runs on a local computer without requiring an internet connection, thus maintaining the security and privacy of the participant information.

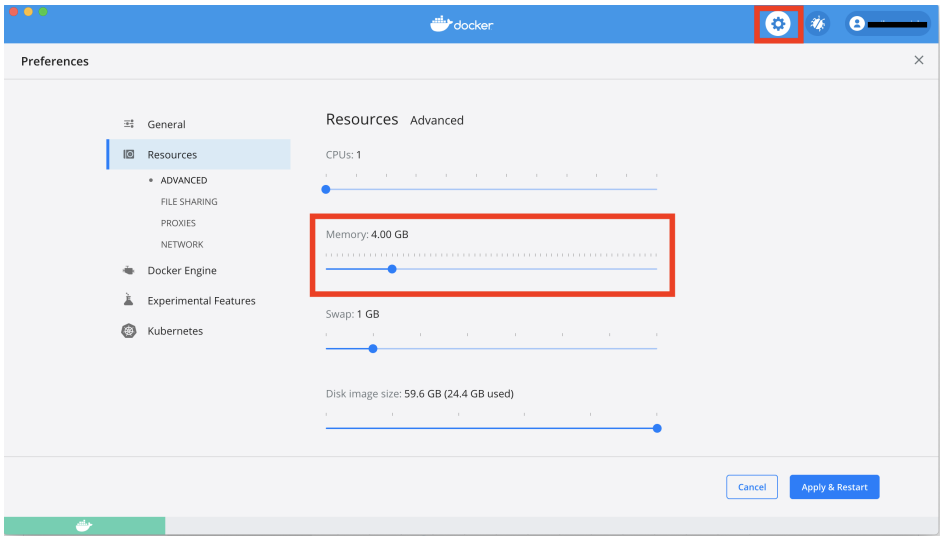
Requirements

- Operating System:
  - MacOS
  - Windows
- RAM: 8GB
- Disk Space: 20GB (docker container is 10GB)
- administrator privileges (initially only, to install the 'docker' software)

Step 0: Install Docker

See the Installing Docker webpage. Make sure you have the latest Docker version installed on your computer.

Note about Docker Settings:
After installing Docker, but before running containers, go to **Docker Settings > Advanced** and change **memory** to greater than 4000 MB (or 4 GiB)



If you are using a Windows computer, also set **CPUs** to 1.

Step 1: Running the STS container

The command to process it through the STS container is:

- macOS:

```
docker run --rm -v $PWD:/tmp ghcr.io/pcgcid/sts_processor:latest \
[--input-file <filename>] [--key-file <keyfile>]
```

- Windows (CMD):

```
docker run --rm -v "%cd%":/tmp ghcr.io/pcgcid/sts_processor:latest ^
[--input-file <filename>] [--key-file <keyfile>]
```

For example, the following command can be used to trip PHI from STS data file and remove procedures occurring in unconsented participants > 18 years of age an STS data stored in 'De-ID STS data fields_NJO.txt' and remove unwanted columns (e.g., those with PHI or other sensitive data) for a list of cases stored in 'sts_pcgc_id_key.txt':

- macOS:

```
docker run --rm -v $PWD:/tmp ghcr.io/pcgcid/sts_processor:latest \
--input-file "De-ID STS data fields_NJO.txt" --key-file "sts_pcgc_id_key.txt"
```

- Windows (CMD):

```
docker run --rm -v "%cd%":/tmp ghcr.io/pcgcid/sts_processor:latest ^
--input-file "De-ID STS data fields_NJO.txt" --key-file "sts_pcgc_id_key.txt"
```

The container will output 'STS_file_for_ACC.tsv' that contains de-identified STS data. The container will also output 'unmapped_mrns.csv' and 'unmapped_sts_ids.csv' files that contain the MRNs and STS IDs that were not found in the key file.

Notes:: - program assumes sts file is a tab delimited text (.txt) file (or .tsv file), that all sites will have the same set of column headers, and that the headers/column names are case sensitive - program assumes a key file is provided that contains the columns MRN, STS_ID, PCGC_ID, reconsented_at_18y - program assumes MRNs and STS IDs in key file map to the format in the STS dataset. For example, if your STS file contains MRNs submitted with various formats such as "MR123456", "mr123456", "123456" then your key should contain MRNs in these various formats

The output files will be stored in the current directory.

Parameters

Command line parameters to show help:

- `-h` or `--help`: Show available parameters. For example, users can use this command:

  ```
  docker run  ghcr.io/pcgcid/sts_processor:latest -h
  ```

  or

  ```
  docker run  ghcr.io/pcgcid/sts_processor:latest --help
  ```

This container **requires** both of the following arguments:

- `--input-file` to specify a tab delimited text file (or .tsv file) with STS data containing STS data. The program assumes that all sites will have the same set of column headers, and that the headers/column names are case sensitive
- `--key-file` to specifya tab delimited text file (or .tsv file) that contains the columns MRN, STS_ID, PCGC_ID, reconsented_at_18y

Details on the processing steps contained in the software

- program will first try to match MRNs in the key file to the MRNs in the sts data file. If a MRN is missing in the key file (e.g., left blank) then the program will try to match record on the STS ID
- program assumes that all patients that were reconsented at >= 18 years of age will be identified by a numeric value of 1 in the reconsented_at_18y column in the key file. For patients < 18 years of age or those that have not yet reconsented after turning 18y this column can either be left blank or a value of 0 provided.