STS extraction

September 6, 2024

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

Michael Wagner Division of Biomedical Informatics Cincinnati Children's Hospital Medical Center September, 2022

Hanh Trang Do Division of Behavioral Med-Clin Psychology Cincinnati Children's Hospital Medical Center August, 2024

based on a previous version of code by

Jason Homsy, M.D., Ph.D. and Marko Boskovski, M.D. Seidman Lab, Dept. of Genetics, Harvard Medical School

Introduction

The STS docker container is used to create a subset of an STS data and remove unwanted columns (e.g., those with PHI or other sensitive data) for a list of cases.

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

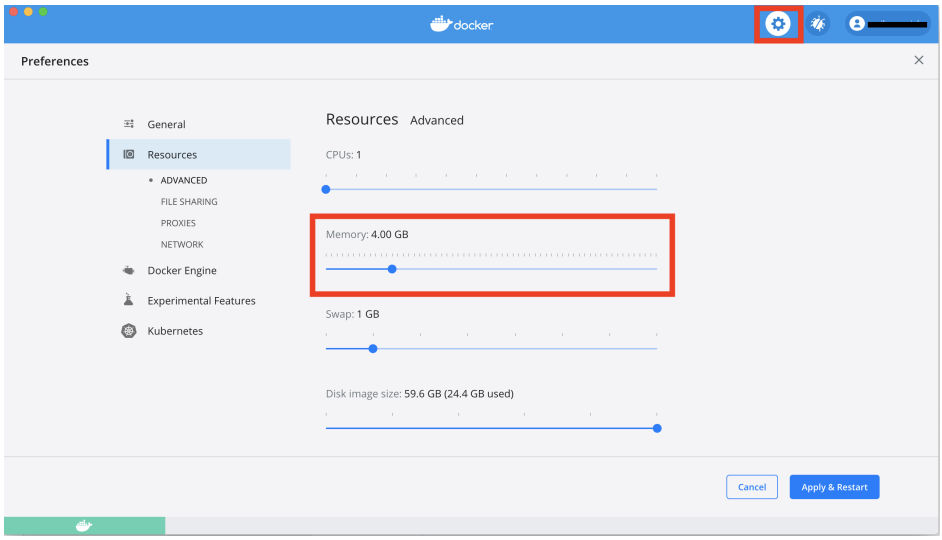
Requirements

- 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.

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:latest \
[--data <data.file>] [--cases <case.file>] [--remove <columns-remove>] \
[-h | --help] [--zip-output-tables]
```

- Windows (CMD):

```
docker run --rm -v "%cd%":/tmp ghcr.io/pcgcid/sts:latest ^
[--data <data.file>] [--cases <case.file>] [--remove <columns-remove>] ^
[-h | --help] [--zip-output-tables]
```

For example, the following command can be used to create a subset of an STS data stored in 'STS.datafile.txt' and remove unwanted columns (e.g., those with PHI or other sensitive data) for a list of cases stored in 'list.of.cases.txt':

- macOS:

```
docker run --rm -v $PWD:/tmp ghcr.io/pcgcid/sts:latest \
--data STS.datafile.txt --cases list.of.cases.txt --remove "PHI1,PHI2"
```

- Windows (CMD):

```
docker run --rm -v "%cd%":/tmp ghcr.io/pcgcid/sts:latest ^
--data STS.datafile.txt --cases list.of.cases.txt
```

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:latest -h
```

or

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

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

- `--data` to specify a flat '|'-separated text file of the site's STS data containing concatenated tables, each table separated by a table name of the format ***tablename
- `--cases` to specify a flat, tab-delimited file with exactly two columns, MEDRECN PCGC.BLINDED.ID

This container takes the following optional arguments:

- `--remove` (optional flag) to remove columns from the output file. The columns to remove (in addition to default columns) should be listed in a comma-separated list following the –remove flag. Default columns to remove:
  - MEDRECN, PATFNAME, PATID, PATLNAME, PATMNAME,PATPOSTALCODE, PATREGION, BIRTHCIT, BIRTHSTA, HOSPNAME, HOSPNPI, HOSPID, HOSPZIP, HOSPSTAT, SURGEON, SURGEONID, SURGNPI, TIN, ASSTSURGEON, ASSTSURGNPI, ASSTSURGEONID,HICNUMBER, PATMINIT, PATCOUNTRY,MATNAMEKNOWN, MATSSNKNOWN, MATLNAME, MATFNAME, MATMINIT, MATMNAME, MATSSN, PARTICID, VENDORID, CNSLTATTND, CNSLTATTNDID, ATTENDSURG, SURGEON,SURGEONID, SURGNPI, ASSTSURGEON, ASSTSURGEONID, ASSTSURGNPI, RESIDENT, RESIDENTID, HOSPZIP, HOSPNPI, REFCARD, REFPHYS, HANDOFFANESTH, HANDOFFSURG, HANDOFFPHYSSTAFF, HANDOFFNURSING, PRIMANESNAME, PRIMANESNPI, SECANES, SECANESNAME, CRNA, CRNANAME, NONCVPHYS, FELRES

  For example, to remove columns named 'PHI1' and 'PHI2' in addition to default columns, use the following syntax: –remove "PHI1,PHI2"

- `--zip-output-tables` (optional flag) to have tab delimited output files zipped into a file called STS_tables.zip (Mac or Linux only)

Running the R script

- This R script can be used to create a subset of an STS data submission and remove unwanted columns (e.g., those with PHI or other sensitive data). The basic syntax is as follows:

```
Rscript filter-STS.R [--data <data.file>] [--cases <case.file>] [--remove <columns-remove>] [-h | -
```

- For example, the following command can be used to create a subset of an STS data stored in 'STS.datafile.txt' and remove unwanted columns (e.g., those with PHI or other sensitive data) for a list of cases stored in 'list.of.cases.txt':

```
Rscript filter-STS.R  --data STS.datafile.txt --cases list.of.cases.txt [--zip-output-tables -help
```

The arguments are the same as for the Docker container. For detailed instructions, see Parameters.

For R script, run script, install packages if missing. Run again. Examine output (check for PHI in output and do other sanity checks.