



Data Preparation – Mac and Linux OS

Data uploads to the AMC Data Commons utilize Amazon Web Services (AWS) S3 storage. Below are instructions for compressing and encrypting data files prior to uploading to the AMC Data Commons AWS S3 storage.

For instructions on Windows systems, please reference the other system-specific instructions documents. If you have any questions or require assistance, email amcdc_data@emmes.com for support.

Before completing data preparation per these instructions, ensure:

A) data submitted are free of PHI/PII

- a. for images, please ensure files and file headers only use Participant ID or Specimen ID, per the Imaging – Pathology Data Standardization Plan.**
- b. for sequencing, use GlobalTrace specimen IDs to identify samples and ensure sequencing files are encrypted prior to upload per instructions below.**

B) an additional backup of all original files is maintained.

1. File Preparation

Confirm that the files to be transferred are saved in the appropriate format. CSV files should be saved as standard Comma Delimited files (not CSV UTF-8). To save an Excel spreadsheet as a CSV file, navigate to Save As, and revise the File Type from Excel Workbook (.xlsx) to CSV (Comma delimited) (*.csv).

2. SHA256 Sums

Note: The following commands should be executed within the Terminal application (/Applications/Utilities/Terminal).

Calculate SHA256 sums for each original uncompressed file to facilitate confirmation of file integrity after the file has been uploaded to the S3 bucket, decrypted, and decompressed. The following command calculates the SHA256 sum for a single file and writes it to a text file:

```
sha256sum <filename> > <filename>_sha256sum.txt
```

The following command calculates the SHA256 sum for a collection of files with a given extension and writes them to a text file:

```
sha256sum *.<extension> > output_sha256sums.txt
```

3. Compression

Compress the files using the gzip command. The following command compresses a single file and retains the original uncompressed version:

```
gzip -k <filename>
```

To compress all the files within a given directory, use the following command:

```
gzip -r <directory>
```

4. Encryption

Encrypt the compressed files using openssl. A password that will be used for encryption & decryption will be needed. ODMC staff will share the encryption password separately. The following command encrypts a single file:

```
openssl enc -aes-256-cbc -md sha256 -pass pass:<password> -in <filename>.gz  
-out <filename>.gz.enc
```

Please use the .enc file extension postfix for all encrypted files. Please add the .enc extension to any existing extension, e.g. for gzipped files, please use <filename>.gz.enc