

Data Management Plan[\[1134\]](#)

1 Introduction[\[1135\]](#)

The Data Management Document covers how the Data was handled during its ML lifecycle. it consists of two Sections "The Data Specification" and "The Data Curation".

The Data Specification plan lists down WHAT data is used in the project. Data Specifications are quantitative (tables, graphs, classes etc.) in nature and provides a quick high level overview of the properties of data used in the project..

The Data Curation describes the tasks which the data went through for curing the data for example Raw Data -> Transform1-> Transform2-> Annotation-> Final versioned data.

2 Data Specification[\[1170\]](#)

2.1 Data Description Details[\[1136\]](#)

Exploration of Data in RDL

Property	Description
Image Modality	Abdominal CT-Scan
Geographic Region	USA
Study Type	Angio, Urogram, Hematuria, Tumor
Data Location	Radiology Data Lake
Studies	157

The filter used was the following:

Filter	Value
Modality	CT
Body Part Examined	ABD
Source (RDL)	Radimetrics
Study Description	Angio, Urogram, Hematuria, Tumor

The exact query used was the following and we randomly selected "n" number of studies for each query:

- Modality: CT **AND** Body Part Examined: ABD **AND** Source (RDL): Radimetrics **AND** Study Description: angio (50 studies)
- Modality: CT **AND** Body Part Examined: ABD **AND** Source (RDL): Radimetrics **AND** Study Description: hematuria (50 studies)
- Modality: CT **AND** Body Part Examined: ABD **AND** Source (RDL): Radimetrics **AND** Study Description: urogram (50 studies)
- Modality: CT **AND** Body Part Examined: ABD **AND** Source (RDL): Radimetrics **AND** Study Description: tumor (7 series)

3 Data Curation[\[1256\]](#)

3.1 Introduction[\[1257\]](#)

The Data Curation plan specifies HOW the data description details listed in the Data Specification section was achieved. The data curation plan is a process and lists down activities inside these processes.

3.2 Discovery and Concept[\[1261\]](#)

Block Diagram / Steps

1. RDL UI Search Filters and Data Transfer to project GCS bucket
2. Data Filtering
3. Additional Data Transfer to project GCS bucket
4. Data Filtering of the combined data
5. Data Conversion to NIFTI
6. GCS versioning for the resulting data
7. Documentation and population details of the data

Data Curation Steps	Description	Link
RDL Search Filters	Modality: CT AND Body Part Examined: ABD AND Source (RDL): Radimetrics AND Study Description: angio (50 studies) This yields us around 158+ Studies.	Link to relevant work item
Data Filtering	Filtering data based on study descriptions and visualizations in order to eliminate bad data	Link to relevant work item
Additional Data Transfer to project GCS bucket	Based on the results of the filtering, data from RDL was again transferred to the project bucket in order to have an appropriate distribution of data. The same filters were used for the data transfer. This transfer included an additional 140 studies which were added to the previous 158 studies, taking the total to 298 studies.	Link to relevant work item Link to relevant work item
Data Filtering of the combined data	The same filtering steps were applied for the combined data from the two transfers and bad data was eliminated and the final number of data points came out to be 428 unique DICOM series.	Link to relevant work item

Data Conversion to NIFTI	Converted the filtered 428 unique DICOM series to NIFTI using a Dataflow script which resulted in 427 NIFTI files.	Link to relevant work item
GCS versioning for the resulting data	Versioned the converted NIFTI data using GCS versioning	Link to relevant work item
Documentation and population details of the data	Making a Table which contains the numbers of samples we have. grouped by Manufacturer of the CT scanner and Patient Sex	Link to relevant work item

https://dev.azure.com/BayerFLATSandbox/Renal%20Stone%20%28REST%29/_apps/hub/devtech-mr.iGVSO-OnPremFeatureProject-mrservice11-002.subHubWork-SmartDocs-OnPrem-teamId=a2262089-d14b-4b52-87b1-7d744c2ece97~filePath=/Data%20Specification.smd

4 Data training Plan[\[1388\]](#)

The Training and testing split do not Apply to our Gan training. we use the entire dataset to train our model. Data version zero get converted into TF records with the image data, affine data and machine attributes. This gives us an easy way to implement cloud accelerated training with multiple GPUs. It also speeds up Data loading.

5 Data Summary[\[1389\]](#)

5.1 Data version 0[\[1390\]](#)

Details about Data version 0 can be found [here](#) switch to preview mode to view the full contents