Imports

In [1]:
```
import logging

from IPython.core.display import display, HTML

from scripts.main import *
from scripts.retrieve_data import *
from scripts.extract_data import *

%load_ext autoreload
%reload_ext autoreload
%autoreload 2

# set the width of the notebook
display(HTML("<style>.container { width:95% !important; }</style>"))
```

Run the whole pipeline

In [2]:
```
run()
```

```
2019-11-14 15:44:43,044|run                           : 28|INFO   | Reading configuration
2019-11-14 15:44:43,065|run                           : 35|INFO   | Starting SchedVisu workflow
2019-11-14 15:44:43,066|retrieve_and_save_data_from_PA: 31|INFO   | Retrieving data from PACS
2019-11-14 15:44:43,069|retrieve_and_save_data_from_PA: 43|INFO   | Processing 20191028
2019-11-14 15:44:43,075|retrieve_and_save_single_day_d: 64|INFO   | Skipping   20191028: save file found at "data\2019\2019-10\2019-10-28.pkl",
nothing to do
2019-11-14 15:44:43,076|retrieve_and_save_data_from_PA: 43|INFO   | Processing 20191029
2019-11-14 15:44:43,079|retrieve_and_save_single_day_d: 64|INFO   | Skipping   20191029: save file found at "data\2019\2019-10\2019-10-29.pkl",
nothing to do
2019-11-14 15:44:43,081|retrieve_and_save_data_from_PA: 43|INFO   | Processing 20191030
2019-11-14 15:44:43,084|retrieve_and_save_single_day_d: 64|INFO   | Skipping   20191030: save file found at "data\2019\2019-10\2019-10-30.pkl",
nothing to do
2019-11-14 15:44:43,086|retrieve_and_save_data_from_PA: 43|INFO   | Processing 20191031
2019-11-14 15:44:43,089|retrieve_and_save_single_day_d: 64|INFO   | Skipping   20191031: save file found at "data\2019\2019-10\2019-10-31.pkl",
nothing to do
2019-11-14 15:44:43,089|retrieve_and_save_data_from_PA: 43|INFO   | Processing 20191101
2019-11-14 15:44:43,093|retrieve_and_save_single_day_d: 64|INFO   | Skipping   20191101: save file found at "data\2019\2019-11\2019-11-01.pkl",
nothing to do
2019-11-14 15:44:43,094|extract_transform_and_save_dat: 18|INFO   | Extracting data from files
2019-11-14 15:44:43,097|extract_transform_and_save_dat: 39|INFO   | Skipping   20191028: save file found at "data\2019\2019-10\2019-10-28.pkl",
loading data
2019-11-14 15:44:43,122|extract_transform_and_save_dat: 39|INFO   | Skipping   20191029: save file found at "data\2019\2019-10\2019-10-29.pkl",
loading data
2019-11-14 15:44:43,142|extract_transform_and_save_dat: 39|INFO   | Skipping   20191030: save file found at "data\2019\2019-10\2019-10-30.pkl",
loading data
2019-11-14 15:44:43,172|extract_transform_and_save_dat: 39|INFO   | Skipping   20191031: save file found at "data\2019\2019-10\2019-10-31.pkl",
loading data
2019-11-14 15:44:43,201|extract_transform_and_save_dat: 39|INFO   | Skipping   20191101: save file found at "data\2019\2019-11\2019-11-01.pkl",
loading data
2019-11-14 15:44:43,231|run                           : 39|INFO   | Finished running SchedVisu workflow
```

Load the configuration

In [20]:
```
config = load_config()
```

Get the series

In [42]:
```
df = extract_transform_and_save_data_from_files(config)
```

```
2019-11-14 16:01:28,688|extract_transform_and_save_dat: 18|INFO    | Extracting data from files
2019-11-14 16:01:28,699|extract_transform_and_save_dat: 39|INFO    | Skipping   20191028: save file found at "data\2019\2019-10\2019-10-28.pkl",
loading data
2019-11-14 16:01:28,796|extract_transform_and_save_dat: 39|INFO    | Skipping   20191029: save file found at "data\2019\2019-10\2019-10-29.pkl",
loading data
2019-11-14 16:01:28,839|extract_transform_and_save_dat: 39|INFO    | Skipping   20191030: save file found at "data\2019\2019-10\2019-10-30.pkl",
loading data
2019-11-14 16:01:28,884|extract_transform_and_save_dat: 39|INFO    | Skipping   20191031: save file found at "data\2019\2019-10\2019-10-31.pkl",
loading data
2019-11-14 16:01:28,927|extract_transform_and_save_dat: 39|INFO    | Skipping   20191101: save file found at "data\2019\2019-11\2019-11-01.pkl",
loading data
```

Do the grouping by

In [39]:
```
df_count_series.loc['PET GE', 'discovery690'].index
```

Out[39]:
```
MultiIndex([('PET CT  CHUV', 'Discovery 690', 'CT'),
            ( 'PET CT CHUV', 'Discovery 690', 'PT')],
           names=['Institution Name', 'Machine', 'Modality'])
```

In [45]:
```
df, df_count_series, df_count_studies = do_series_groupby(config, df)
display(df_count_series)
display(df_count_studies)
display(df_count_series.groupby('Machine Group').sum())
display(df_count_studies.groupby('Machine Group').sum())
```

					Number of Series
Machine Group	**Machine Group List**	**Institution Name**	**Machine**	**Modality**	
Discovery	brightspeed,tandemdiscovery670	CHUV	BrightSpeed	CT	49
			Tandem_Discovery_670	NM	55
Intevo	encore2,symbiaintevo16	Radiologie CHUV	Encore2	NM	68
				PT	18
			Symbia Intevo 16	CT	50
Millennium	millenniummpr	MEDECINE NUCLEAIRE CHUV LAUSANNE	MILLENNIUM MPR	NM	83
PET GE	discovery690	PET CT CHUV	Discovery 690	CT	135
				PT	160
PET Siemens	biograph64,biograph64vision600	Centre hospitalier universitaire vaudois	Biograph64	CT	180
			Biograph64_Vision 600	PT	232
mixed cases	biograph64	Centre hospitalier universitaire vaudois	Biograph64	CT	3
	biograph64vision600	Centre hospitalier universitaire vaudois	Biograph64_Vision 600	PT	49
	brightspeed	CHUV	BrightSpeed	CT	2
	brightspeed,millenniummpr,tandemdiscovery670	CHUV	BrightSpeed	CT	2
			Tandem_Discovery_670	NM	1
		MEDECINE NUCLEAIRE CHUV LAUSANNE	MILLENNIUM MPR	NM	4
	encore2	Radiologie CHUV	Encore2	NM	8
				PT	4
	symbiaintevo16	Radiologie CHUV	Symbia Intevo 16	CT	1
	tandemdiscovery670	CHUV	Tandem_Discovery_670	NM	22

Machine Group	Machine Group List	Institution Name	Machine	Modality	Number of Studies
Discovery	brightspeed,tandemdiscovery670	CHUV	BrightSpeed	CT	13
			Tandem_Discovery_670	NM	13
Intevo	encore2,symbiaintevo16	Radiologie CHUV	Encore2	NM	15
				PT	10
			Symbia Intevo 16	CT	16
Millennium	millenniummpr	MEDECINE NUCLEAIRE CHUV LAUSANNE	MILLENNIUM MPR	NM	13
PET GE	discovery690	PET CT CHUV	Discovery 690	CT	43
				PT	49
PET Siemens	biograph64,biograph64vision600	Centre hospitalier universitaire vaudois	Biograph64	CT	43
			Biograph64_Vision 600	PT	43
mixed cases	biograph64	Centre hospitalier universitaire vaudois	Biograph64	CT	2
	biograph64vision600	Centre hospitalier universitaire vaudois	Biograph64_Vision 600	PT	7
	brightspeed	CHUV	BrightSpeed	CT	1
	brightspeed,millenniummpr,tandemdiscovery670	CHUV	BrightSpeed	CT	1
			Tandem_Discovery_670	NM	1
		MEDECINE NUCLEAIRE CHUV LAUSANNE	MILLENNIUM MPR	NM	1
	encore2	Radiologie CHUV	Encore2	NM	5
				PT	3
	symbiaintevo16	Radiologie CHUV	Symbia Intevo 16	CT	1
	tandemdiscovery670	CHUV	Tandem_Discovery_670	NM	9

Machine Group	Number of Series
Discovery	104
Intevo	136
Millennium	83
PET GE	295
PET Siemens	412
mixed cases	96

Machine Group	Number of Studies
Discovery	26
Intevo	41
Millennium	13
PET GE	92
PET Siemens	86
mixed cases	31

Clean up the data (while changing things in the API so the bugs/errors do not appear anymore)

In [44]:
```
df = df[~df['Study Instance UID'].isin(df[df['Institution Name'] == 'Hopital neuchatelois']['Study Instance UID'])]
df = df[~df['Study Instance UID'].isin(df[df['Institution Name'] == 'MEDECINE NUCLEAIRE']['Study Instance UID'])]
df = df[~df['Study Instance UID'].isin(df[df['Machine'] == '']['Study Instance UID'])]
```

Check why some non-valid institution name went through the filter

In [9]:
```
df_bad_series = df[df['Study Instance UID'].isin(df[df['Machine'] == 'Ingenuity TF PET/CT']['Study Instance UID'])]

if len(df_bad_series) > 0:
    bad_study_UID = list(set(df_bad_series['Study Instance UID'].values))[0]
    logging.info('bad_study_UID: ' + bad_study_UID)

    inst_name = list(set([inst_name.replace('  ', ' ') for inst_name in df_bad_series.loc[:, 'Institution Name']]))[0]
    logging.info('inst_name: ' + inst_name)
    logging.info('accepted_inst_names: ' + str(accepted_inst_names))
    logging.info('inst_name is in accepted_inst_names? ' + str(inst_name.lower().replace(' ', '') in accepted_inst_names))
    logging.info('date: ' + str(list(set(df_bad_series['Series Date']))[0]))

    df_studies = find_studies_for_day(config, '20191029', ['PT', 'NM'])
    df_bad_study = df_studies[df_studies['Study Instance UID'] == bad_study_UID]
    df_series = find_series_for_studies(config, df_bad_study)
    df_series
```

Solve the problem of having some studies with mixed machine names

In [11]:
```
# find a study which has both the millennium and another machine
df_series_for_study = df[df['Study Instance UID'].isin(df[df['Machine'] == 'MILLENNIUM MPR']['Study Instance UID'])]
df_series_for_study[df_series_for_study['Study Instance UID'].isin(df_series_for_study[df_series_for_study['Machine'] == 'BrightSpeed']['Study Instance UID'])]
```

Out[11]:

	Study Instance UID	Machine Group List	Series Date	Series Time	Modality	Institution Name	Study Description	Series Description	Patie
1124	1.2.826.0.1.3680043.2.146.2.20.3293563.1900233...	brightspeed,millenniummpr,tandemdiscovery670	20191031	151206	NM	CHUV	PTHY PINHOLE	STAT 5min	32935
1125	1.2.826.0.1.3680043.2.146.2.20.3293563.1900233...	brightspeed,millenniummpr,tandemdiscovery670	20191031	153633	NM	MEDECINE NUCLEAIRE CHUV LAUSANNE	PTHY PINHOLE	PTHY 45MIN 5CM	32935
1126	1.2.826.0.1.3680043.2.146.2.20.3293563.1900233...	brightspeed,millenniummpr,tandemdiscovery670	20191031	155133	NM	MEDECINE NUCLEAIRE CHUV LAUSANNE	PTHY PINHOLE	PTHY 60MIN 5C-1	32935
1127	1.2.826.0.1.3680043.2.146.2.20.3293563.1900233...	brightspeed,millenniummpr,tandemdiscovery670	20191031	155254	CT	CHUV	PTHY PINHOLE	CTRecon0scoutSPECT CT	32935
1128	1.2.826.0.1.3680043.2.146.2.20.3293563.1900233...	brightspeed,millenniummpr,tandemdiscovery670	20191031	155445	CT	CHUV	PTHY PINHOLE	CT tete cou NUC std	32935
1129	1.2.826.0.1.3680043.2.146.2.20.3293563.1900233...	brightspeed,millenniummpr,tandemdiscovery670	20191031	170724	NM	MEDECINE NUCLEAIRE CHUV LAUSANNE	PTHY PINHOLE	PTHY 2H15 5CM	32935
1130	1.2.826.0.1.3680043.2.146.2.20.3293563.1900233...	brightspeed,millenniummpr,tandemdiscovery670	20191031	172934	NM	MEDECINE NUCLEAIRE CHUV LAUSANNE	PTHY PINHOLE	THYR 10MIN 5CM	32935

In [ ]: