

PIC-SURE

NIH Cloud Platform Interoperability

March 17th 2020

Arnaud Serret-Larmande, MD Msc
Avillach Lab, Department of Biomedical Informatics
Harvard Medical School



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE overview



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE overview

- Part of the [BioData Catalyst](#) initiative
- Studies and variables exploration tool + data retrieving



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE overview

- Clinical and genomic datasets integration from multiple TOPMed and TOPMed related studies
- Two components: Graphical User Interface and API
- Graphical User Interface: Allows investigator to search available data and conduct feasibility queries, allowing for cohorts to be built in real-time and results to be exported
- Cohort building and analysis directly through the PIC-SURE API.



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE components



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

Platform components

Graphical User Interface:

- Access phenotypic variables (and soon genomic ones)
- Search all phenotypic data by typing terms directly in the search bar, across all available studies
- Search available harmonized variables across multiple studies
- Develop and refine cohorts in real-time by building complex queries using multiple variables
- Refine queries and subset cohort based on variable values
- Export full phenotypic datasets to analysis environment with the click of a button

R and Python API

- Explore studies and variables
- Query data directly from the API
- Learn how to use the API through [public example Jupyter notebooks](#), available on GitHub and cloud computing platforms (Seven Bridges and Terra)
- Save query identifiers to easily locate past queries in the future



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE specificity

- Under appropriate governance, can allow users to search aggregate data and see what variables and cohorts are available prior to creating an account or officially requesting data
- Develop and refine their cohort using clinical (and soon genomic) data
- Direct access to harmonized variables
- Data download from Graphical User Interface, or data query from the API
- Easy access to the data of interest: for example, the whole Framingham Heart Study can be retrieved in one tabular format
- PIC-SURE API in two languages (R and python) and works from multiple cloud-computing platforms using Jupyter notebooks or RStudio Server.



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

User Interface

<https://picsure.biodatacatalyst.nhlbi.nih.gov/>

- Controlled access, login via eRA Commons

The screenshot displays the BioData CATALYST web application interface. At the top, the NIH logo is followed by the text "National Heart, Lung, and Blood Institute". To the right, the "BioData CATALYST" logo is prominently displayed, with "Powered by PIC-SURE" underneath it. A navigation bar contains four buttons: "Query Builder" (which is highlighted), "User Profile", "Help", and "Log Out".

The main content area is titled "QUERY BUILDER". On the left, there is a search bar with a magnifying glass icon and the placeholder text "Search...". On the right, a large light blue box features the number "241956" in a large, bold font, with the text "Total Participants" below it. At the bottom of this box is a button labeled "Select Data For Export".

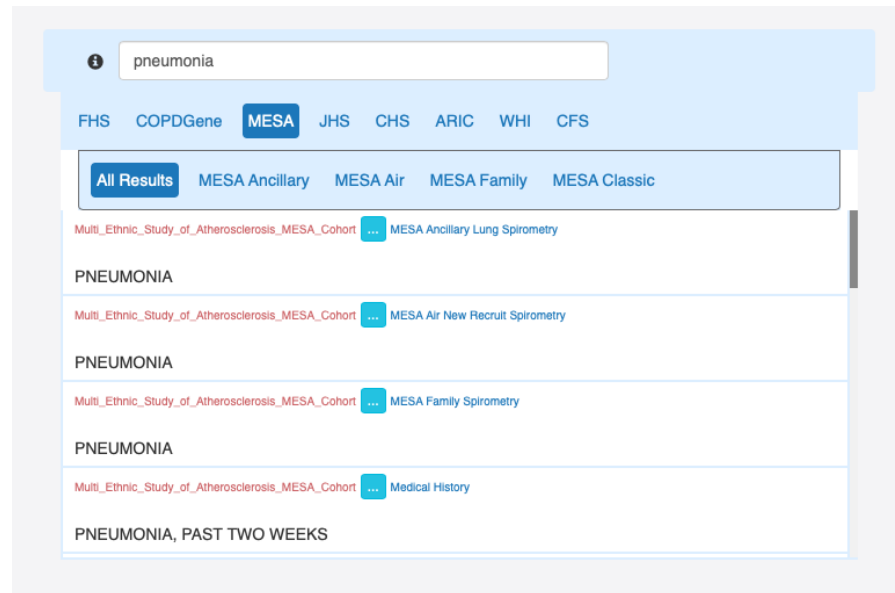


HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

User Interface

- Variables from 29 studies
 - 241,956 patients
 - 43 harmonized variables
-
- Variable selection through search engine



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

User Interface

- Study feasibility through cross-counts using variable values
- Possible data export after cohort constitution

QUERY BUILDER

Genetic_Epidemiology_of_COPD_COPDGene_ back delete edit

Age at enrollment, Greater than or equal to 70

value range, min: 39.9 - max: 85 Run

By numeric value Greater than or equal to 70

AND

Genetic_Epidemiology_of_COPD_COPDGene_ back delete edit

Have you ever smoked cigarettes?, Yes

Restrict By Value Run

Available Values	Selected Values
Filter...	Yes
No	
Select All »	« Select None

1579
Total Participants

Select Data For Export

data

Prepare Data Export



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE API

- Available in two languages: R and python
- Access data using cloud environments
 - Authentication through user specific token
 - Remote access, no need to download data locally
- Enable to access 20+ different studies through the same API



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE API

Three different example notebooks publicly available in both R and python:

- PIC-SURE 101
- Phenome-Wide analysis example
- Harmonized variables analysis examples

GitHub repo: https://github.com/hms-dbmi/Access-to-Data-using-PIC-SURE-API/tree/master/NHLBI_BioData_Catalyst

File Edit View Run Kernel Git Tabs Settings Help

/ ... / NHLBI_BioData_Catalyst / python /

Name	Last Modified
python_lib	6 hours ago
get_your_token.ipynb	6 hours ago
HarmonizedVariables_analysis.ipynb	5 hours ago
PheWAS.ipynb	4 hours ago
PICSURE_API_101.ipynb	4 hours ago
requirements.txt	7 hours ago
token.txt	6 hours ago

Launcher

HarmonizedVariables_anal

PICSURE_API_101.ipynb

```
mask_count = variablesDict["observationCount"].between(100, 2000)
varnames = variablesDict.loc[mask_cat & mask_count, "name"]

[25]: my_query.filter().add(yo_stop_smoking_varname, min=20, max=70)
      my_query.select().add(varnames[:50])

[25]: <PicSureHpdsLib.PicSureHpdsAttrListKeys.AttrListKeys at 0x7fd589e9bdd8>
```

Retrieving the data

Once our query object is finally built, we use the `query.run()` function to retrieve the data corresponding to our query

```
[26]: query_result = my_query.getResultsDataFrame(low_memory=False)

[27]: query_result.shape

[27]: (4678, 52)

[28]: query_result.head()

[28]:
```

Patient ID	Exam	Cohort	Phenotype	Access and Quality of Care	Observations
ARIC_70AAs_4032	Visit 5	Physical	Observations	Difficulty in Obtaining Care	Q11a. C. Difficulty in obtaining care. In the past 12 months unable to afford being seen by doctor
ARIC_70AAs_4032	Visit 5	Physical	Observations	Difficulty in Obtaining Care	Q11b. C. Difficulty in obtaining care. In the past 12 months unable to afford mental health care
ARIC_70AAs_4032	Visit 5	Physical	Observations	Difficulty in Obtaining Care	Q11c. C. Difficulty in obtaining care. In the past 12 months unable to afford nursing home care
ARIC_70AAs_4032	Visit 5	Physical	Observations	Difficulty in Obtaining Care	Q11d. C. Difficulty in obtaining care. In the past 12 months unable to afford nursing home care



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE API: Cross studies variables information



```
[22]: dictionary = resource.dictionary()
dictionary_search = dictionary.find("COPD")
```

```
[23]: dictionary_search.DataFrame().head()
```

```
[23]:
```

	min	categorical	patientCount	observationCount	max
KEY					
\Genetic Epidemiology of COPD (COPDGene)\Subject Phenotype\CT Acquisition Parameters\CT Slicer\Percent gas trapping total lung: CT Slicer\	0.0214	False	8276	8276	87.8387
\Genetic Epidemiology of COPD (COPDGene)\Subject Phenotype\Respiratory Disease\Family History\Asthma: Father or Mother\Father: asthma\	NaN	True	10098	10098	NaN
\Genetic Epidemiology of COPD (COPDGene)\Subject Phenotype\Respiratory Disease\Respiratory Conditions\Hayfever\Hayfever: diagnosed by doctor or other health professional\	NaN	True	3000	3000	NaN



```
[10]: dictionary_search <- hpds::find.in.dictionary(resource, "COPD")
```

```
plain_variablesDict <- hpds::find.in.dictionary(resource, "COPDGene") %>%
hpds::extract.dataframe()
```

```
[11]: plain_variablesDict[10:20,]
```

	name	min	categorical	patientCount	observationCount	max
10	\Genetic Epidemiology of COPD (COPDGene)\Subject Phenotype\Medical History\Cardio Vascular Diseases\Heart attack [MI]\	NA	TRUE	10099	10099	NA
11	\Genetic Epidemiology of COPD (COPDGene)\Subject Phenotype\Respiratory Disease\Respiratory Symptoms\Shortness of Breath\Too breathless to leave the house on dressing/undressing\	NA	TRUE	6418	6418	NA
12	\Genetic Epidemiology of COPD (COPDGene)\Subject Phenotype\SF-36 Health Survey\SF-36 General Health (GHE) <score/normalized>\	16.23	FALSE	4573	4573	63.9

PIC-SURE API: Data querying



```
[90]: my_query.require().add(smoke)
      my_query.filter().add(stroke, values="Yes")
      my_query.select().add(varnames)
      query_result = my_query.getResultsDataFrame()
```



```
[52]: hpds::query.require.add(my_query, keys = smoke)
      hpds::query.filter.add(my_query,
                             keys = stroke,
                             values="Yes")
      hpds::query.select.add(my_query, keys = varnames)
      my_df <- hpds::query.run(my_query, result.type = "dataframe")
```

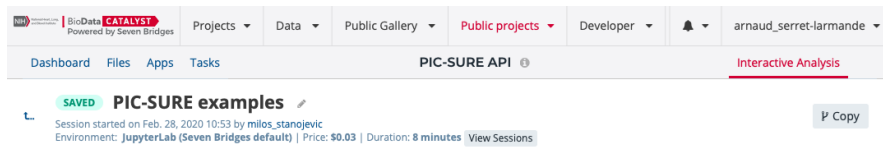


HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

Publicly available Workspaces

Seven Bridges



Files Settings

Analysis files:

- README.md
- python
 - requirements.txt
 - get_your_token.ipynb
 - PheWAS.ipynb
 - PICSURE_API_101.ipynb
 - HarmonizedVariables_analysis.ipynb
- python_lib
- R

Produced by this analysis

No files

PICSURE python API use-case: Phenome-Wide analysis on BioData Catalyst studies

This notebook is an illustration example of how to use the python **PIC-SURE API** to select and query data from a database (HPDS format). It takes as use-case a simple PheWAS analysis. This notebook is intentionally straightforward, and explanation provided are only aimed at guiding through the PheWAS analysis pipeline. For a more step-by-step introduction to the python PIC-SURE API, see the [PICSURE-API_101.ipynb](#) notebook.

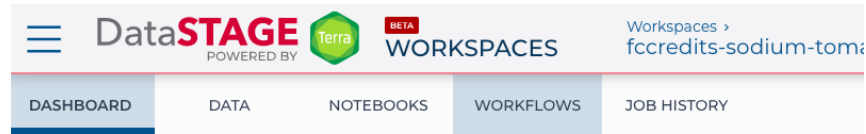
Before running this notebook, please be sure to get a user-specific security token. For more information on how to proceed, see the [get_your_token.ipynb](#) notebook

Environment set-up

System requirements

- Python 3.6 or later

Terra (available soon)



ABOUT THE WORKSPACE

Python PIC-SURE_API BioData Catalyst examples

Jupyter Notebooks examples of PIC-SURE API use-cases, using BioData Catalyst studies. PIC-SURE API is available in two languages: R and python. Python PIC-SURE API requires python 3.6 or later.

Before running the example notebooks, first-time user must go through two first steps, detailed below:

1. Getting user personal security token to get access to the BioDataCatalyst Data
2. Setting-up your Terra Workspace

1. Getting user personal security token to get access to the BioDataCatalyst Data



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

Policy



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

Policy

- Use of BioData Catalyst policies:
 - Currently, a user can perform search with authentication only (no authorization) and get back aggregate counts for phenotypic study variables.
 - Counts resulting in less than 10 will be displayed as <10 (for example, a query resulting in 5 participants will be displayed as less than 10)
 - Counts resulting in zero will be displayed as 0
 - Stigmatizing variable search not allowed
 - No genomic data search without authorization
- Direct download of certain types of data are not allowed



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS

PIC-SURE

Links:

- [BioData Catalyst front-page](#)
- [PIC-SURE User Interface](#)
- [GitHub public repo: Jupyter and RMarkdown notebooks example](#)
- [Seven Bridges publicly available Workspaces](#)



HARVARD
MEDICAL SCHOOL

BLAVATNIK INSTITUTE
BIOMEDICAL INFORMATICS