Building a Data Catalog

Promoting Data Reuse and Collaboration at an Academic Medical Center

Kevin Read, MLIS, MAS
Alisa Surkis, PhD, MLIS
### AMA Physician Masterfile

**Alternate Title(s)**: American Medical Association, Physician Masterfile, AMA Masterfile

**Description**: The Physician Masterfile includes education, training, and professional certification information on virtually all Doctors of Medicine (MD) and Doctors of Osteopathic Medicine (DO) in the United States, Puerto Rico, Virgin Islands, and certain Pacific Islands. It includes current and historical data for more than 1.4 million physicians, residents, and medical students in the United States. The file includes approximately 411,000 graduates of foreign medical schools who reside in the United States and who have met the educational and credentialing requirements necessary for recognition.

**Subject**: Delivery of Health Care, Population Characteristics

**Geographic Coverage**: National

**Timeline**: 1935 - present

**Access Rights**: Fee required

**Local Expert**: Heather Gold

### American Community Survey

**Alternate Title(s)**: ACS

**Description**: The American Community Survey (ACS) is an ongoing survey that provides data every year, giving communities the current information they need to plan investments and services. Information from the survey generates data that help determine how more than $600 billion in federal and state funds are distributed each year. To help communities, state governments, and federal programs, we ask about age, sex, race, family and relatedship, income and benefits, health insurance, education, veteran status, disabilities, where you work and how far you get there, and where you live and how much you pay for some essentials.
### External Datasets

**AMA Physician Masterfile**
- **Alternate Title(s)**: American Medical Association Physician Masterfile, AMA Masterfile
- **Description**: The Physician Masterfile includes education, training, and professional certification information on virtually all Doctors of Medicine (MD) and Doctors of Osteopathic Medicine (DO) in the United States, Puerto Rico, Virgin Islands, and certain Pacific Islands. It includes current and historical data for more than 1.4 million physicians, residents, and medical students in the United States. The file includes approximately 411,000 graduates of foreign medical schools who reside in the United States and who have met the educational and credentialing requirements necessary for recognition.
- **Subject**: Delivery of Health Care
- **Geographic Coverage**: National
- **Timeframe**: 1935 - present
- **Access Rights**: Fee required
- **Local Expert**: Heather Gold

**American Community Survey**
- **Alternate Title(s)**: ACS
- **Description**: The American Community Survey (ACS) is an ongoing survey that provides data every year — giving communities the current information they need to plan investments and services. Information from the survey generates data that help determine how more than $600 billion in federal and state funds are distributed each year. To help communities, state governments, and federal programs, we ask about age, sex, race, family and relationship, income and benefits, health insurance, education, veteran status, disability, where you work and how you get there, and whether you live and how much you pay for some essentials.
- **Geographic Coverage**: National, Multiple States

### Internal Datasets

- **Subject Domain**
  - Delivery of Health Care
  - Population Characteristics
  - Health Care Costs
  - Health Care Utilization
  - Health Status
  - Quality of Health Care
  - Risk Factors
  - Chronic Disease
  - Cancer
  - Timeframe
    - 2000-2009
    - 2010-2019
    - 1980-1989
    - Prior to 1980
  - Geographic Coverage
    - National
    - Multiple States
NYU Data Catalog

THE INSTITUTIONAL PLAYERS
Clinical and Translational Science Institute (CTSI)

• Bridging gaps across different schools at NYU and beyond

• Mission of promoting collaboration

• Role in data sharing
Department of Population Health (DPH)

- Research using existing large datasets
- Willingness to share data
- Clinical informatics hub
Research IT

• Provides applications for clinical research data collection

• Research data storage

• Data governance
DataCore

NYULMC DataCore

Clinical Research Data Management
- Single Center Studies
- Multicenter Studies

Accessing Clinical Data for Research; Data Warehouse
- Clinical Data
- Claims Data
- Omics Data

Data Set Curation
- External Data Sets
- Internal Data Sets
Libraries

NYU Health Sciences Library
• Works mainly with researchers at NYU School of Medicine, College of Dentistry

NYU Division of Libraries
• Serve Global Public Health program, School of Nursing, Bioengineering, Neuroscience, Psychology, Nutrition, etc.
Libraries

NYU Health Sciences Library
• Works mainly with researchers at NYU School of Medicine, College of Dentistry

NYU Division of Libraries
• Serve Global Public Health program, School of Nursing, Bioengineering, Neuroscience, Psychology, Nutrition, etc.

NYU Abu Dhabi Library
• Increasing amount of research at campus
NYU Data Catalog

REASONS FOR IMPLEMENTATION
Heavy users of large, external datasets (e.g. census, national health surveys, Medicare)
Heavy users of large, external datasets (e.g. census, national health surveys, Medicare)

Lack of knowledge about institutional licenses

Department of Population Health
Difficulty accessing datasets

Heavy users of large, external datasets (e.g. census, national health surveys, Medicare)

Lack of knowledge about institutional licenses

Difficulty accessing datasets

Difficulty working with datasets

Department of Population Health
NYU Data Catalog

STRATEGY PART 1:
OUR INITIAL APPROACH
Give the people what they want

• Focus on external datasets first
• Deal with internal datasets in the future
• Add functionality wherever possible
Local experts

- Identified researchers through grant and publication data
- Worked with DPH for outreach
- Includes researchers (and staff) from NYU and NYULMC
- Facilitates research and fosters collaboration
Facilitate access

- Host NY Statewide Planning and Research Cooperative System (SPARCS) data
- University Health Consortium (UHC) data
Other functionality

- PubMed searches

- Related datasets
  - Linkage datasets (e.g. MEPS and NHIS)
  - Datasets from same publisher (e.g. HCUP)
  - National and local dataset versions (e.g. NHANES and NYCHANES)
# NIH Data Sharing Repositories

This table lists NIH-supported data repositories that accept submissions of appropriate data from NIH-funded investigators (and others). Also included are resources that aggregate information about biomedical data and information sharing systems. The table can be sorted according by name and by NIH Institute or Center and may be searched using keywords so that you can find repositories more relevant to your data. Links are provided to information about submitting data to and accessing data from the listed repositories. Additional information about the repositories and points-of-contact for further information or inquiries can be found on the websites of the individual repositories.

<table>
<thead>
<tr>
<th>IC</th>
<th>Repository Name</th>
<th>Repository Description</th>
<th>Data Submission Policy</th>
<th>Access to Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCI</td>
<td>The Cancer Imaging Archive (TCIA)</td>
<td>The Cancer Imaging Archive (TCIA) is a large archive of medical images of cancer accessible for public download. All images are stored in DICOM file format. The images are organized as 'collections', typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc.) or research focus.</td>
<td>How to Submit Data to TCIA</td>
<td>How to Access TCIA Data</td>
</tr>
<tr>
<td>NCI (NHGRI, NIGMS)</td>
<td>PeptideAtlas</td>
<td>PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences.</td>
<td>How to Submit Data to PeptideAtlas</td>
<td>How to Access PeptideAtlas Data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>FlyBase: A Drosophila Genomic and Genetic Database</td>
<td>Drosophila Genomic and Genetic database that includes proteomics data, microarrays and Tiling BACs.</td>
<td>How to Submit Data to FlyBase</td>
<td>How to Access FlyBase Data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>The Zebrafish Model Organism Database (ZFIN)</td>
<td>ZFIN serves as the zebrafish model organism database. It aims to: a) be the community database resource for the laboratory use of zebrafish, b) develop and support integrated zebrafish genetic, genomic and developmental information, c) maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e) facilitate the use of zebrafish as a model for human biology, and f) serve the needs of the research community.</td>
<td>How to Submit Data to ZFIN</td>
<td>How to Access ZFIN Data</td>
</tr>
<tr>
<td>NHGRI</td>
<td>WormBase</td>
<td>WormBase is an international consortium of biologists and computer scientists dedicated to providing the research community with accurate, current, accessible information concerning the genetics, genomics and biology of C. elegans and related nematodes.</td>
<td>How to Submit Data to WormBase</td>
<td>How to Access WormBase Data</td>
</tr>
</tbody>
</table>
Existing metadata schemas

- W3C Data Catalog Vocabulary
- Dryad Metadata Schema
- DataCite Metadata Schema
- NPG’s Scientific Data
Strategy over purity

• Need to describe both external and internal datasets
• PubMed: search link vs article link
• Local Experts = external data only
• Authorship
# Metadata

<table>
<thead>
<tr>
<th>Publisher URL</th>
<th>hyperlink to publisher</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access path URL</td>
<td>hyperlink to dataset. Multiple</td>
<td>x</td>
</tr>
<tr>
<td>Associated Publication DOI</td>
<td>DOI</td>
<td>x</td>
</tr>
<tr>
<td>Associated Publication Citation</td>
<td>PubMed format citation or Citing Medicine format [<a href="http://www.ncbi.nlm.nih.gov/books/NBK7266/">http://www.ncbi.nlm.nih.gov/books/NBK7266/</a>]</td>
<td>x</td>
</tr>
<tr>
<td>PubMed Search</td>
<td>hyperlink to PubMed search</td>
<td>x</td>
</tr>
<tr>
<td>Dataset format</td>
<td>ASCII, multiple fields available.</td>
<td>x</td>
</tr>
<tr>
<td>Dataset size</td>
<td>MIME type. Multiple</td>
<td>x</td>
</tr>
<tr>
<td>Subject:Geographic coverage</td>
<td>temporary categories for display in population health</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>National</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Multiple States</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>New York (State)</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>New York (City)</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>International</td>
<td>x</td>
</tr>
<tr>
<td>Subject:Geographic coverage detail</td>
<td>Foundation Center Classification (e.g., for New York City = North America - New York (State) - New York). Multiple</td>
<td>x</td>
</tr>
<tr>
<td>Subject: Start Date</td>
<td>Dataset coverage start date YYYY-MM-DD</td>
<td>x</td>
</tr>
<tr>
<td>Subject: End Date</td>
<td>Dataset coverage end date YYYY-MM-DD</td>
<td>x</td>
</tr>
<tr>
<td>Subject:Timeframe</td>
<td>Simplify interface for Cathy - Above fields will ultimately replace this field. Multiple</td>
<td>x</td>
</tr>
<tr>
<td>Subject:Domain</td>
<td>MeSH -- with LC as an optional extension. Multiple.</td>
<td>x</td>
</tr>
<tr>
<td>Health Care Costs</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Quality of Health Care</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Delivery of Health Care</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Population Characteristics</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Health Status</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Risk Factors</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Cancer</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Health Care Utilization</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
NYU Data Catalog

THE DATA CATALOG
# The Data Catalog

**AMA Physician Masterfile**

<table>
<thead>
<tr>
<th>Alternate Title(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Medical Association Physician Masterfile, AMA Masterfile</td>
<td></td>
</tr>
</tbody>
</table>

- **Subject**: Delivery of Health Care
- **Population Characteristics**
- **Health Care Costs**
- **Health Care Utilization**
- **Health Status**
- **Quality of Health Care**
- **Risk Factors**
- **Chronic Disease**
- **Cancer**

**Geographic Coverage**: National

**Timeframe**: 1900 - present

**Access Rights**: Fee required

**Local Expert**: Heather Gold

**American Community Survey**

<table>
<thead>
<tr>
<th>Alternate Title(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACS</td>
<td></td>
</tr>
</tbody>
</table>

- **Geographic Coverage**: National, Multiple States

Information from the survey generates data that help determine how more than $400 billion in federal and state funds are distributed each year. To help communities, state governments, and Federal programs, we ask about: age, sex, race, family and relationships, income and benefits, health insurance, education, veteran status, disabilities, where you work and how you get there, and where you live and how much you pay for some essentials.

[http://datacatalog.med.nyu.edu/](http://datacatalog.med.nyu.edu/)
Current population health focus

- Maintaining functionality as catalog expands in scope
- Increase in subject domains
Centers for Medicare and Medicaid Services

Description
Centers for Medicare and Medicaid Services (CMS) offers researchers and other health care professionals a broad range of quantitative information on programs, from estimates of future Medicare and Medicaid spending to enrollment, spending, and claims data, and a broad range of consumer research to help its partners and staff. CMS also conducts demonstration projects to explore alternative policies of health care coverage and delivery. National health expenditure data not specific to the Medicare or Medicaid programs are also included.

Publisher
United States, Centers for Medicare and Medicaid Services

Timeframe
2002 - present

Geographic Coverage
United States

Local Expert
Caroline Baum
Saul Blecker
Heather Gold
Keith Goldfield
Danil Makarov
Tod Mjanovitch
James Slover

Subject Matter
Health Care Costs
Delivery of Health Care

Keywords
Medicare
Medicaid
Health expenditures

Access URL
Link

Access Restrictions
Free to all
Fee/Application required

Access Instructions
The CMS DataNavigator provides users with the ability to search CMS data by program, setting/type of care, topic, geography and document type. Some of the CMS data available through the DataNavigator is freely available to download from the website. Each freely available dataset has an associated policy statement, support contact information, and downloadable ZIP files. Multiple dataset descriptions can also be batch downloaded into excel files. For restricted use datasets, users must fill out a CMS Access Request Form in order to gain access to the data and understand the privacy and security restrictions when using these restricted CMS data files. Restricted or limited use datasets are also a payed service and requests will be submitted through pay.gov.

Dataset Type
Administrative Data

Dataset Format(s)
PDF, RTF, Microsoft Word, Microsoft Excel, CSV, XML

Dataset Size
15-36 MB

PubMed Search
Articles using this dataset

Related Datasets
Surveillance, Epidemiology, and End Results Program

Related Dataset - Note
SEER-Medicare linked database available here: http://appliedresearch.cancer.gov/seeermedicare/
Contact Us

Use this form to contact our team with any questions, comments, or suggestions. We look forward to hearing from you!

KiID/NetID/Bellevue ID

First Name

Last Name

Email

Affiliation: NYU School of Medicine

Reason for Contacting Us:
- Volunteer as a local expert
- Suggest an external dataset
- Request uploading of your dataset(s)
- General inquiry or comments

Please give more details about your question or comment

Submit
Technical Components

• Apache Solr search engine

• Javascript using:
  • Backbone.js* (structure)
  • jQuery (webpage interaction)
  • Underscore.js (display)

• Will use Google Analytics
ISSUES: THE BIG THREE
PROMOTION

NO USE

NO RESOURCES

SCALABILITY

FEW DATASET ENTRIES

MAINTENANCE

POORLY MAINTAINED DATA
Promotion: Consumers and producers

• Combination of reaching out to targeted groups and broad email blasts

• Many owners increases visibility
  • DataCore, CTSI, Library, DPH

• Data discoverable through multiple systems
  • NYU Main Library Catalog, Profiles RNS
Scalability: Factors and issues

• Curation time per dataset/researcher
• Number of FTEs
• Researcher upload of datasets
Maintenance: People and data in motion

- Researcher information
  - Feed from Faculty Data Mart
  - HR feeds from main campus

- People that leave vs data that leaves
  - Published vs. archival records

- ORCID for all researchers by Spring 2015

- Tracking data location
BONUS ISSUE
How do we define a dataset?

 DATA SET

Scientific Paper

Comparison of radiotherapy alone with radiotherapy plus hyperthermia in locally advanced pelvic tumours: a prospective randomised, multicentre trial.

van der Zee J, Gonzalez Gonzalez D, van Rhoon GC, van Dijk JD, van Putten WL, Hort AA. Subdivision of Hyperthermia, Academic Medical Centre, Amsterdam, The Netherlands

Background:
Local-control rates after radiotherapy for locally advanced tumours of the bladder, cervix, and rectum are disappointing. We investigated the effect of adding hyperthermia to standard radiotherapy.

Methods:
The study was a prospective, randomised, multicentre trial. 338 patients were enrolled from 1990 to 1996, in cancer centres in the Netherlands, who had bladder cancer stages T2, T3, or T4, or cervical cancer stages IB, IIa, or III, or IV, or rectal cancer stage MO-1 were assessed. Patients were randomly assigned radiotherapy (median total dose 65 Gy) alone (n=176) or radiotherapy plus hyperthermia (n=182). Our primary endpoints were complete response and duration of local control. We did the analysis by intention to treat.

Findings:
Complete-response rates were 39% after radiotherapy and 55% after radiotherapy plus hyperthermia (p=0.001). The duration of local control was significantly longer with radiotherapy plus hyperthermia than with radiotherapy alone (p=0.04). Treatment effect did not differ significantly by tumour site, but the addition of hyperthermia seemed to be most important for cervical cancer, for which the complete-response rate with radiotherapy plus hyperthermia was 83% compared with 57% after radiotherapy alone (p=0.005). 3-year overall survival was 27% in the radiotherapy group and 51% in the radiotherapy plus hyperthermia group, for bladder cancer, an initial difference in local control disappeared during follow-up.

Interpretation:
Hyperthermia in addition to standard radiotherapy may be especially useful in locally advanced cervical tumours. Studies of larger numbers of patients are needed for other pelvic tumour sites before practical recommendations can be made.

Published:
Cancer 2000 Apr 1;95(5):1119-25

Overall survival data from Bolla trial

Radiotherapy with hormones
Radiotherapy (no hormones)
NYU Data Catalog

STRATEGY PART 2: NEXT STEPS
Building a community of users

• Strategy over purity
  • External datasets form a strong core

• Initial internal datasets will be within similar domains to external datasets
  • DPH

• Aiming for scope creep
  • Clinical research data
  • Social science data
  • Basic science data down the road…
Internal data curation

• Low bar for inclusion
  • Data on a flash drive? Okay…for now

• Ingest slowly
  • Addition of internal datasets may lead to changes in workflow and metadata
Heavy users of large, external datasets (e.g. census, national health surveys, Medicare)

Lack of knowledge about institutional licenses

Difficulty accessing datasets

Difficulty working with datasets

Willingness to share research data

Department of Population Health
Promotional strategies

- Demo the data catalog at DPH faculty meeting – hours ago!
- Reach out to small research areas in DPH
- Demo to Global Institute of Public Health
  - 11 schools across NYU
NYU Data Catalog

FUTURE DIRECTIONS
NYULMC DataCore

Clinical Research Data Management
- Single Center Studies
- Multicenter Studies

Accessing Clinical Data for Research; Data Warehouse
- Clinical Data
- Claims Data
- Omics Data

Data Set Curation
- External Data Sets
- Internal Data Sets
Clinical Research Data Flow
Hosted Internal & External Data Flow (Future State)

Data Flow Business Owners/Collaborators:
- NYU Health Sciences Library
- Clinical Translational Science Institute
- MCIT/Research IT/DataCore
Cross-campus collaboration

• Pushing data from catalog into NYU Library Main catalog

• Outreach from main campus librarians for ingesting internal datasets
Integrating with Profiles

Eric Meeks

Title: Technical Lead, CTSI
School: UCSF School of Medicine
Department: CTSI
Address: 550 16th Street, 6th Floor
San Francisco, CA 94158
Phone: 415-514-0081
Email: eric.meeks@ucsf.edu

Awards and Honors
University of California 2013 University of California Larry L. Sautter Award - Golden Award

Overview
Eric Meeks is the Technical Architect for the CTSI Virtual Home team at UCSF. He is driving the effort to bring the OpenSocial standard into research networking tools and has worked extensively with Harvard Profiles and the OpenSocial foundation to achieve this end.

In industry, Mr. Meeks has worked extensively in Silicon Valley with numerous entrepreneurial efforts supporting first-generation products and services, including ping, one of the first social network systems to support OpenSocial.

Eric is co-inventor for two patents resulting from his work with early web transaction systems. He earned an undergraduate degree in engineering from Duke University and studied computational fluid dynamics at Princeton prior to embarking on a career in software development.

Interests
Research Networking

Websites
- OpenSocial Article
- Work on Github
- Profile on LinkedIn

In The News
- Eric Meeks to Lead OpenSocial Gadget Modernization Project (June 2, 2014)

Many gadgets are already available
We've developed a library of open source gadgets you can install on your research networking system. Most of these gadgets will work in VIVO or Profiles without alteration. Others have back end processes that are specifically coded to the Profiles data structure, and a few are specific to UCSF data sources. However, all are open source and we hope that you will use these to make your particular research network sites more feature rich, making alterations as necessary and then contributing back to the ORNG open source community.

Featured publications

Emphasize highly cited or most relevant publications.
- Developer: UCSF
- Works on VIVO and Profiles
- View source code
- See an example

Videos

Incorporate a YouTube viewer into profile pages so owners can add video lectures and other presentations to their profile. Videos are hosted in the cloud, via the free YouTube service.
- Developer: UCSF
- Works on VIVO and Profiles
- View source code
- See an example

NIH grant search

Match profile owners to top grant opportunities from NIH RePORTER.
- Developer: Wake Forest
Continuing the conversation
Thank you!

kevin.read@nyumc.org
alisa.surkis@nyumc.org