MEDICAL DATA FOR MACHINE LEARNING

This is a curated list of medical data for machine learning.
This list is provided for informational purposes only, please make sure you respect any and all usage restrictions for any of the data listed here.

1. MEDICAL IMAGING DATA

EchoNet-Dynamic
A Large New Cardiac Motion Video Data Resource for Medical Machine Learning, from Stanford. Overview: https://echonet.github.io/dynamic/index.html Access: https://echonet.github.io/dynamic/index.html#access

---

The National Library of Medicine presents MedPix®
Database of 53,000 medical images from 13,000 patients with annotations. **Requires registration**.
Information: https://medpix.nlm.nih.gov/home

ABIDE: The Autism Brain Imaging Data Exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism.
Function MRI images for 539 individuals suffering from ASD and 573 typical controls. These 1112 datasets are composed of structural and resting state functional MRI data along with an extensive array of phenotypic information. **Requires registration**.
Paper: http://www.ncbi.nlm.nih.gov/pubmed/23774715
Information: http://fcon_1000.projects.nitrc.org/indi/abide/
Preprocessed version: http://preprocessed-connectomes-project.org/abide/

Alzheimer's Disease Neuroimaging Initiative (ADNI)
MRI database on Alzheimer's patients and healthy controls. Also has clinical, genomic, and biomaker data. **Requires registration**.
Paper: http://www.neurology.org/content/74/3/201.short
Access: http://adni.loni.usc.edu/data-samples/access-data/

CT Colongraphy for Colon Cancer (Cancer Imaging Archive) CT scan for diagnosing of colon cancer. Includes data for patients without polyps, 6-9mm polyps, and greater than 10 mm polyps. Access: https://wiki.cancerimagingarchive.net/display/Public/CT+COLONOGRAPHY#dc149b9170f54aa29e88f1119e25ba3e

---

Digital Retinal Images for Vessel Extraction (DRIVE)
The DRIVE database is for comparative studies on segmentation of blood vessels in retinal images. It consists of 40 photographs out of which 7 showing signs of mild early diabetic retinopathy.
Paper: https://ieeexplore.ieee.org/document/1282003
Access: http://www.isi.uu.nl/Research/Databases/DRIVE/download.php

---

AMRG Cardiac Atlas The AMRG Cardiac MRI Atlas is a complete labelled MRI image set of a normal patient's heart acquired with the Auckland MRI Research Group 's Siemens Avanto scanner. The atlas aims to provide university and school students, MR technologists, clinicians...

Congenital Heart Disease (CHD) Atlas The Congenital Heart Disease (CHD) Atlas represents MRI data sets, physiologic clinical data and computer models from adults and children with various congenital heart defects. The data have been acquired from several clinical centers including Rady...

DETERMINE Defibrillators to Reduce Risk by Magnetic Resonance Imaging Evaluation, is a prospective, multicenter, randomized clinical trials in patients with coronary artery diseases and mild-to-moderate left ventricular dysfunction. The primary objective...

MESA Multi-Ethnic Study of Atherosclerosis, is a large-scale cardiovascular population study (>6,500 participants) conducted in six centres in the USA. It aims to investigate the manifestation of subclinical to clinical cardiovascular disease before...

OASIS The Open Access Series of Imaging Studies (OASIS) is a project aimed at making MRI data sets of the brain freely available to the scientific community. Two datasets are available: a cross-sectional and a longitudinal set.

- Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults: This set consists of a cross-sectional collection of 416 subjects aged 18 to 96. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD). Additionally, a reliability data set is included containing 20 nondemented subjects imaged on a subsequent visit within 90 days of their initial session.
- Longitudinal MRI Data in Nondemented and Demented Older Adults: This set consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For

> each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

Access: http://www.oasis-brains.org/

Isic Archive - Melanoma This archive contains 23k images of classified skin lesions. It contains both malignant and benign examples.

Each example contains the image of the lesion, meta data regarding the lesion (including clasisfication and segmentation) and meta data regarding the patient.

The data can be viewed in this link: https://www.isic-archive.com (in the gallery section)
It can be downloaded through the site or by using this repository:
https://github.com/GalAvineri/ISIC-Archive-Downloader

SCMR Consensus Data The SCMR Consensus Dataset is a set of 15 cardiac MRI studies of mixed pathologies (5 healthy, 6 myocardial infarction, 2 heart failure and 2 hypertrophy), which were acquired from different MR machines (4 GE, 5 Siemens, 6 Philips). The main objectives...

Sunnybrook Cardiac Data The Sunnybrook Cardiac Data (SCD), also known as the 2009 Cardiac MR Left Ventricle Segmentation Challenge data, consist of 45 cine-MRI images from a mixed of patients and pathologies: healthy, hypertrophy, heart failure with infarction and heart...

Access: http://www.cardiacatlas.org/studies/

Lung Image Database Consortium (LIDC)

Preliminary clinical studies have shown that spiral CT scanning of the lungs can improve early detection of lung cancer in high-risk individuals. Image processing algorithms have the potential to assist in lesion detection on spiral CT studies, and to assess the stability or change in lesion size on serial CT studies. The use of such computer-assisted algorithms could significantly enhance the sensitivity and specificity of spiral CT lung screening, as well as lower costs by reducing physician time needed for interpretation.

The intent of the Lung Imaging Database Consortium (LIDC) initiative was to support a consortium of institutions to develop consensus guidelines for a spiral CT lung image resource and to construct a database of spiral CT lung images. The investigators funded under this initiative created a set of guidelines and metrics for database use and for developing a database as a test-bed and showcase for those methods. The database is available to researchers and users through the Internet and has wide utility as a research, teaching, and training resource.

Specifically, the LIDC initiative aims were to provide:

- a reference database for the relative evaluation of image processing or CAD algorithms and
- a flexible query system that will provide investigators the opportunity to evaluate a wide range of technical parameters and de-identified clinical information within this database that may be important for research applications.

This resource will stimulate further database development for image processing and CAD evaluation for applications that include cancer screening, diagnosis, and image guided intervention, and treatment. Therefore, the NCI encourages investigator-initiated grant applications that utilize the database in their research. NCI also encourages investigator-initiated grant applications that provide tools or methodology that may improve or complement the mission of the LIDC.

Access: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI#

TCIA Collections

Cancer imaging data sets across various cancer types (e.g. carcinoma, lung cancer, myeloma) and various imaging modalities. The image data in The Cancer Imaging Archive (TCIA) is organized into purpose-built collections of subjects. The subjects typically have a cancer type and/or anatomical site (lung, brain, etc.) in common. Each link in the table below contains information concerning the scientific value of a collection, information about how to obtain any supporting non-image data which may be available, and links to view or download the imaging data. To support reproducibility in scientific research, TCIA supports Digital Object Identifiers (DOIs) which allow users to share subsets of TCIA data referenced in a research manuscript.

Access: http://www.cancerimagingarchive.net/

Belarus tuberculosis portal

Tuberculosis (TB) is a major problem of Belarus Public Health. Recently situation has been complicated with emergence and development of MDR/XDR TB and HIV/TB which require long-term treatment. Many and the most severe cases usually disseminate across the country to different TB dispensaries. The ability of leading Belarus TB specialists to follow such patients will be greatly improved by using a common database containing patients’ radiological images, lab work and clinical data. This will also significantly improve adherence to the treatment protocol and result in a better record of the treatment outcomes. Criteria for inclusion clinical cases in the database of the portal - patients admitted to the MDR-TB department of RSPC of Pulmonology and Tuberculosis with diagnosed or suspected of MDR-TB, which conducted CT – study (± 2 months from the date of registration) Belarus dataset have both chest X-rays and CT scans of the same patient.

Access: http://tuberculosis.by/

DDSM: Digital Database for Screening Mammography

The Digital Database for Screening Mammography (DDSM) is a resource for use by the mammographic image analysis research community. Primary support for this project was a grant from the Breast Cancer Research Program of the U.S. Army Medical Research and Materiel Command. The DDSM project is a collaborative effort involving co-p.i.s at the Massachusetts General Hospital (D. Kopans, R. Moore), the University of South Florida (K. Bowyer), and Sandia National Laboratories (P. Kegelmeyer). Additional cases from Washington University School of Medicine were provided by Peter E. Shile, MD, Assistant Professor of Radiology and Internal Medicine. Additional collaborating institutions include Wake Forest University School of Medicine (Departments of Medical Engineering and Radiology), Sacred Heart Hospital and ISMD, Incorporated. The primary purpose of the database is to facilitate sound research in the development of computer algorithms to aid in screening. Secondary purposes of the database may include the development of algorithms to aid in the diagnosis and the development of teaching or training aids. The database contains approximately 2,500 studies. Each study includes two images of each breast, along with some associated patient information (age at time of study, ACR breast density rating, subtlety rating for abnormalities, ACR keyword description of abnormalities) and image information (scanner, spatial resolution, ...). Images containing suspicious areas have associated pixel-level "ground truth" information about the locations and types of suspicious regions. Also provided are software both for accessing the mammogram and truth images and for calculating performance figures for automated image analysis algorithms.

Access: http://marathon.csee.usf.edu/Mammography/Database.html

---

INbreast: Database for Digital Mammography

The INbreast database is a mammographic database, with images acquired at a Breast Centre, located in a University Hospital (Hospital de São João, Breast Centre, Porto, Portugal). INbreast has a total of 115 cases (410 images) of which 90 cases are from women with both breasts (4 images per case) and 25 cases are from mastectomy patients (2 images per case). Several types of lesions (masses, calcifications, asymmetries, and distortions) are included. Accurate contours made by specialists are also provided in XML format.

Access: http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database

---

mini-MIAS: MIAS MiniMammographic Database

The Mammographic Image Analysis Society (MIAS) is an organisation of UK research groups interested in the understanding of mammograms and has generated a database of digital mammograms. Films taken from the UK National Breast Screening Programme have been digitised to 50 micron pixel edge with a Joyce-Loebl scanning microdensitometer, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. The database contains 322 digitised films and is available on 2.3GB 8mm (ExaByte) tape. It also includes radiologist's "truth"-markings on the locations of any abnormalities that may be present. The database has been reduced to a 200 micron pixel edge and padded/clipped so that all the images are 1024x1024. Mammographic images are available via the Pilot European Image Processing Archive (PEIPA) at the University of Essex.

Access: http://peipa.essex.ac.uk/info/mias.html

Prostate

Prostate cancer (CaP) has been reported on a worldwide scale to be the second most frequently diagnosed cancer of men accounting for 13.6% (Ferlay et al. (2010)). Statistically, in 2008, the number of new diagnosed cases was estimated to be 899,000 with no less than 258,100 deaths (Ferlay et al. (2010)).

Magnetic resonance imaging (MRI) provides imaging techniques allowing to diagnose and localize CaP. The I2CVB provides a multi-parametric MRI dataset to help at the development of computer-aided detection and diagnosis (CAD) system. Access: http://i2cvb.github.io/

Access: http://www.ehealthlab.cs.ucy.ac.cy/index.php/facilities/32-software/218-datasets

- **MRI Lesion Segmentation in Multiple Sclerosis Database**
- **Emergency Tele-Orthopedics X-ray Digital Library**
- **IMT Segmentation**
- **Needle EMG MUAP Time Domain Features**

DICOM image sample sets These datasets are exclusively available for research and teaching. You are not authorized to redistribute or sell them, or use them for commercial purposes.

All these DICOM files are compressed in JPEG2000 transfer syntax.

Access: http://www.osirix-viewer.com/resources/dicom-image-library/

SCR database: Segmentation in Chest Radiographs

The automatic segmentation of anatomical structures in chest radiographs is of great importance for computer-aided diagnosis in these images. The SCR database has been established to facilitate comparative studies on segmentation of the lung fields, the heart and the clavicles in standard posterior-anterior chest radiographs.

In the spirit of cooperative scientific progress, we freely share the SCR database and are committed to maintaining a public repository of results of various algorithms on these segmentation tasks. On these pages, instructions can be found on downloading the database and uploading results, and benchmark results of various methods can be inspected.

Access: http://www.isi.uu.nl/Research/Databases/SCR/

Medical Image Databases & Libraries

Access: http://www.omnimedicalsearch.com/image_databases.html

General Category

- e-Anatomy.org - Interactive Atlas of Anatomy - e-anatomy is an anatomy e-learning web site. More than 1500 slices from normal CT and MR exams were selected in order to cover the entire sectional anatomy of human body. Images were labeled using Terminologia Anatomica. A user-friendly interface allows to cine through multi-slice image series combined with interactive textual information, 3D models and anatomy drawings.
- Medical Pictures and Definitions - Welcome to the largest database of medical pictures and definitions on the Internet. There are many sites sites that provide medical information but very few that provide medical pictures. As far as we know we are the only one that provides a medical picture database with basic information about each term pictured. Editor's Note: Nice website with free access & no pesky registration to 1200+ health and medical related images with definitions.
- Nucleus Medical Art - Medical Illustrations, Medical Art. Includes 3D animations. "Nucleus Medical Art, Inc. is a leading creator and distributor of medical illustrations, medical animations, and interactive multimedia for publishing, legal, healthcare, entertainment, pharmaceutical, medical device, academia and other markets, both in the U.S. and abroad. Editors Note: Great website.
- Medical Image Databases on the Internet (UTHSCSA Library) - A directory of links to websites with topic specific medical related images.
- Surgery Videos - A National Library of Medicine MedlinePlus collection of links to 100s and 100s of different surgical procedures. You must have RealPlayer media player on your computer to view these videos which are free of charge.
- The ADAM Medical Encyclopedia with Illustrations. Perhaps one of the best illustrated medical works on the internet today, the ADAM Medical Encyclopedia includes over 4,000 articles about diseases, tests, symptoms, injuries, and surgeries. It also contains an extensive library of medical photographs and illustrations to back up those 4,000 articles. These illustrations and articles are free to the public.
- Hardin MD - Medical and Disease Pictures, is a Free and established resource that has

been offered by the University of Iowa for quite some time. The home page is in directory style where users will have to drill down to find the images they are looking for, many of which go offsite. Nevertheless, Hardin MD is an excellent gateway to 1,000s of detailed medical photos and illustrations.

- Health Education Assets Library (HEAL) - Health on the Net Foundation Media Gallery Headquartered in Switzerland, (HON) is an international body that seeks to encourage ethical provision of online health information. "HONmedia (the image gallery) is an unique repository of over 6'800 medical images and videos, pertaining to 1,700 topics and themes. This peerless database has been created manually by HON and new image links are constantly being added from the world-wide Web. HON encourages users to make their own image links available via the Submit an image link." Library includes anatomical images, visual affects of diseases and conditions and procedures.
- Public Health Image Library (PHIL) Created by a Working Group at the Centers for Disease Control and Prevention (CDC), the PHIL offers an organized, universal electronic gateway to CDC's pictures. We welcome public health professionals, the media, laboratory scientists, educators, students, and the worldwide public to use this material for reference, teaching, presentation, and public health messages. The content is organized into hierarchical categories of people, places, and science, and is presented as single images, image sets, and multimedia files.
- Images from the History of Medicine - This system provides access to the nearly 60,000 images in the prints and photograph collection of the History of Medicine Division (HMD) of the U.S. National Library of Medicine (NLM). The collection includes portraits, pictures of institutions, caricatures, genre scenes, and graphic art in a variety of media, illustrating the social and historical aspects of medicine.
- Pozemedicale.org - Collection of medical images in Spanish, Italian, Portuguese and Italian.
- Old Medical Pictures: Hundreds of fascinating and interesting old, but high quality photographs and images from the late 19th and early 20th century.

Subject Speciality Image Libraries and Collections

- Anatomy of the Human Body by Henry Gray - The Bartleby.com edition of Gray’s Anatomy of the Human Body features 1,247 vibrant engravings—many in color—from the classic 1918 publication.
- The Crookston Collection - A collection of medical slides taken by Dr. John H. Crookston that have been digitized and are available to the public and doctors.
- DAVE Project - A searchable library of gastrointestinal endoscopic video clips covering a wide spectrum endoscopic imaging.
- Dermnet - Browsable collection of over 8,000 high quality, dermatology images.
- Interactive Dermatology Atlas - Image reference source for common and uncommon skin problems.
- The Multi-Dimensional Human Embryo is a collaboration funded by the National Institute of Child Health and Human Development (NICHD) to produce and make available over the internet a three-dimensional image reference of the Human Embryo based on magnetic resonance imaging.
- GastroLab Endoscopy Archives Was initiated in 1996 with the goal of maintaining an endoscopic image gallery free to use for all interested health care personals.
- MedPix Is a Radiology and Medical Picture Databases resource tool. The home page interface is confusing and the entire website design is not user-friendly and has a mid 1990s feel to it. However, if you have the time (patience) it could prove to be an important resource for some.
- OBGYN.net Image Library - This site is devoted entirely to providing access to images of

> interest to women's health. In addition to providing you with access to OBGYN.net images we also point to other women's health related images on the Internet. Because of the graphic nature of the material some individuals may prefer not to view these images.They are provided for educational purposes only.

VIA Group Public Databases

Documented image databases are essential for the development of quantitative image analysis tools especially for tasks of computer-aided diagnosis (CAD). In collaboration with the I-ELCAP group we have established two public image databases that contain lung CT images in the DICOM format together with documentation of abnormalities by radiologists. Please access the links below for more details:

Access: http://www.via.cornell.edu/databases/

CVonline: Image Databases Access: http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm

---

The USC-SIPI Image Database The USC-SIPI image database is a collection of digitized images. It is maintained primarily to support research in image processing, image analysis, and machine vision. The first edition of the USC-SIPI image database was distributed in 1977 and many new images have been added since then.

The database is divided into volumes based on the basic character of the pictures. Images in each volume are of various sizes such as 256x256 pixels, 512x512 pixels, or 1024x1024 pixels. All images are 8 bits/pixel for black and white images, 24 bits/pixel for color images. The following volumes are currently available:

```
Textures  Brodatz textures, texture mosaics, etc.
Aerials   High altitude aerial images
Miscellaneous   Lena, the mandrill, and other favorites
Sequences   Moving head, fly-overs, moving vehicles
```

Access: http://sipi.usc.edu/database/

Histology dataset: image registration of differently stain slices

The dataset consists of 2D histological microscopy tissue slices, stained with different stains, and landmarks denoting key-points in each slice. The task is image registration - align all slices in particular set of images (consecutive stain cuts) together, for instance to the initial image plane. The main challenges for these images are the following: very large image size, appearance differences, and lack of distinctive appearance objects. The dataset contains 108 image pairs and manually placed landmarks for registration quality evaluation.

Access: http://cmp.felk.cvut.cz/~borovji3/?page=dataset

2. CHALLENGES/CONTEST DATA

Visual Concept Extraction Challenge in Radiology Manually annotated radiological data of several anatomical structures (e.g. kidney, lung, bladder, etc.) from several different imaging modalities (e.g. CT and MR). They also provide a cloud computing instance that anyone can use to develop and evaluate models against benchmarks.

Access: http://www.visceral.eu/

Grand Challenges in Biomedical Image Analysis

A collection of biomedical imaging challenges in order to *facilitate better comparisons between new and existing solutions*, by standardizing evaluation criteria. You can create your own challenge as well. As of this writing, there are 92 challenges that provide downloadable data sets.

Access: http://www.grand-challenge.org/

Dream Challenges

DREAM Challenges pose fundamental questions about systems biology and translational medicine. Designed and run by a community of researchers from a variety of organizations, our challenges invite participants to propose solutions — fostering collaboration and building communities in the process. Expertise and institutional support are provided by Sage Bionetworks, along with the infrastructure to host challenges via their Synapse platform. Together, we share a vision allowing individuals and groups to collaborate openly so that the "wisdom of the crowd" provides the greatest impact on science and human health.

- The Digital Mammography DREAM Challenge.
- ICGC-TCGA DREAM Somatic Mutation Calling RNA Challenge (SMC-RNA)
- DREAM Idea Challenge
- These were the active challenges at the time of adding, many more past challenges and upcoming challenges are present!

Access: http://dreamchallenges.org/

Kaggle diabetic retinopathy

High-resolution retinal images that are annotated on a 0–4 severity scale by clinicians, for the detection of diabetic retinopathy. This data set is part of a completed Kaggle competition, which is generally a great source for publicly available data sets.

Access: https://www.kaggle.com/c/diabetic-retinopathy-detection

---

Cervical Cancer Screening

In this kaggle competition, you will develop algorithms to correctly classify cervix types based on cervical images. These different types of cervix in our data set are all considered normal (not cancerous), but since the transformation zones aren't always visible, some of the patients require further testing while some don't.

Access: https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data

---

Multiple sclerosis lesion segmentation

challenge 2008. A collection of brain MRI scans to detect MS lesions.

Access: http://www.ia.unc.edu/MSseg/

Multimodal Brain Tumor Segmentation Challenge

Large data set of brain tumor magnetic resonance scans. They've been extending this data set and challenge each year since 2012.

Access: http://braintumorsegmentation.org/

Coding4Cancer

A new initiative by the Foundation for the National Institutes of Health and Sage Bionetworks to host a series of challenges to improve cancer screening. The first is for digital mammography readings. The second is for lung cancer detection. The challenges are not yet launched.

Access: http://coding4cancer.org/

EEG Challenge Datasets on Kaggle

- Melbourne University AES/MathWorks/NIH Seizure Prediction - Predict seizures in long-term human intracranial EEG recordings

Access: https://www.kaggle.com/c/melbourne-university-seizure-prediction

- American Epilepsy Society Seizure Prediction Challenge - Predict seizures in intracranial EEG recordings

Access: https://www.kaggle.com/c/seizure-prediction

- UPenn and Mayo Clinic's Seizure Detection Challenge - Detect seizures in intracranial

EEG recordings

Access: https://www.kaggle.com/c/seizure-detection

- Grasp-and-Lift EEG Detection - Identify hand motions from EEG recordings

Access: https://www.kaggle.com/c/grasp-and-lift-eeg-detection

---

Challenges track in MICCAI Conference

The Medical Image Computing and Computer Assisted Intervention. Most of the challenges would've been covered by websites like grand-challenges etc. You can still see all of them under the "Satellite Events" tab of the conference sites.

- 2019 - https://www.miccai2019.org/programme/workshops-challenges-tutorials/#tablepress-10
- 2018 - https://www.miccai2018.org/en/WORKSHOP---CHALLENGE---TUTORIAL.html
- 2017 - http://www.miccai2017.org/satellite-events
- 2016 - http://www.miccai2016.org/en/SATELLITE-EVENTS.html
- 2015 - https://www.miccai2015.org/frontend/index.php?page_id=589

Access: http://www.miccai.org/ConferenceHistory

International Symposium on Biomedical Imaging (ISBI)

The IEEE International Symposium on Biomedical Imaging (ISBI) is a scientific conference dedicated to mathematical, algorithmic, and computational aspects of biomedical imaging, across all scales of observation. Most of these challenges will be listed in grand-challenges. You can still access it by visiting the "Challenges" tab under "Program" in each year's website.

- 2019 - https://biomedicalimaging.org/2019/challenges/
- 2018 - https://biomedicalimaging.org/2018/challenges/
- 2017 - http://biomedicalimaging.org/2017/challenges/
- 2016 - http://biomedicalimaging.org/2016/?page_id=416

Access: http://biomedicalimaging.org

Continuous Registration Challenge (CRC)

Continuous Registration Challenge (CRC) is a challenge for registration of lung- and brain images inspired by modern software development practices. Participants implement their algorithm using the open source SuperElastix C++ API. The challenge focuses on pairwise registration of lungs and brains, two problems frequently encountered in clinical settings. They have collected seven open-access data sets and one private data set (3+1 lung data sets, 4 brain data sets). The challenge results will be presented and discussed at the upcoming Workshop On Biomedical Image Registration (WBIR 2018).

Access: https://continuousregistration.grand-challenge.org/home/

Automatic Non-rigid Histological Image Registration (ANHIR)

This ANHIR challenge aims at the automatic nonlinear image registration of 2D whole slice imaging (WSI) microscopy images of histopathology tissue samples stained with different dyes. The task is difficult due to non-linear deformations affecting the tissue samples, different appearance of each stain, repetitive texture, and the large size of the whole slide images.

- Challenge: https://anhir.grand-challenge.org/
- Benchmark: http://borda.github.io/BIRL
- Refernce: BIRL: Benchmark on Image Registration methods with Landmark validation

Bone X-Ray Deep Learning Competition using MURA

MURA (musculoskeletal radiographs) is a large dataset of bone X-rays. The Stanford ML Group and AIMI Center are hosting a competition where algorithms are tasked with determining whether an X-ray study is normal or abnormal. The algorithms are evaluated on a test set of 207 musculoskeletal studies, where each study was individually retrospectively labeled as normal or abnormal by 6 board-certified radiologists. Three of these radiologists were used to create a gold standard, defined as the majority vote of the labels of the radiologists, and the other three were used to obtain the best radiologist performance, defined as the maximum score of the three radiologists with the gold standard as groundtruth. The challenge leaderboard is hosted publicly and updated every two weeks.

Access: https://stanfordmlgroup.github.io/competitions/mura/

2019 Kidney and Kidney Tumor Segmentation Challenge (KiTS19)

The KiTS19 challenge is on the semantic segmentation of kidneys and kidney tumors in contrast-enhanced CT scans. The dataset consists of 300 patients with preoperative arterial-phase abdominal CTs annotated by experts. 210 (70%) of these were released as a training set and the remaining 90 (30%) were held out as a test set. This challenge was held in conjunction with MICCAI 2019.

Access: https://github.com/neheller/kits19/

3. DATA DERIVED FROM ELECTRONIC HEALTH RECORDS (EHRS)

Building the graph of medicine from millions of clinical narratives

Co-occurence statistics for medical terms extracted from 14 million clinical notes and 260,000 patients.

Paper: http://www.nature.com/articles/sdata201432

Data: http://datadryad.org/resource/doi:10.5061/dryad.jp917

Learning Low-Dimensional Representations of Medical Concept

Low-dimensional embeddings of medical concepts constructed using claims data. Note that this paper utilizes data from *Building the graph of medicine from millions of clinical narratives*

Paper: http://cs.nyu.edu/~dsontag/papers/ChoiChiuSontag_AMIA_CRI16.pdf

Data: https://github.com/clinicalml/embeddings

MIMIC-III, a freely accessible critical care database

Anonymized critical care EHR database on 38,597 patients and 53,423 ICU admissions. **Requires registration**.

Paper: http://www.nature.com/articles/sdata201635

Data: http://physionet.org/physiobank/database/mimic3cdb/

Clinical Concept Embeddings Learned from Massive Sources of Medical Data

Embeddings for 108,477 medical concepts learned from 60 million patients, 1.7 million journal articles, and clinical notes of 20 million patients

Paper: https://arxiv.org/abs/1804.01486

Embeddings: https://figshare.com/s/00d69861786cd0156d81

Interactive tool: http://cui2vec.dbmi.hms.harvard.edu

Evaluation of Embeddings of Laboratory Test Codes for Patients at a Cancer Center

200 dimensional Word2Vec embeddings of 1098 laboratory test codes (LOINCs) trained from 8,280,238 lab orders for 79,081 patients at City of Hope National Medical Center (Los Angeles, CA).

Paper: https://arxiv.org/abs/1907.09600

Embeddings and Code: https://github.com/elleros/DSHealth2019_loinc_embeddings

---

4. NATIONAL HEALTHCARE DATA

Centers for Disease Control and Prevention (CDC)

Data from the CDC on many areas, including:

- Biomonitoring
- Child Vaccinations
- Flu Vaccinations

- Health Statistics
- Injury & Violence
- MMWR
- Motor Vehicle
- NCHS
- NNDSS
- Pregnancy & Vaccination
- STDs
- Smoking & Tobacco Use
- Teen Vaccinations
- Traumatic Brain Injury
- Vaccinations
- Web Metrics

Landing page: https://data.cdc.gov
Data Catalog: https://data.cdc.gov/browse

Medicare Data
Data from the Centers for Medicare & Medicaid Services (CMS) on hospitals, nursing homes, physicians, home healthcare, dialysis, and device providers.
Landing page: https://data.medicare.gov
Explorer: https://data.medicare.gov/data

Texas Public Use Inpatient Data File Data on 11 Million inpatient visits with diagnosis, procedure codes and outcomes from Texas between 2006 & 2009.

Link: https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm

---

Dollars for Doctors
Propublica investigation of money paid by pharmaceutical companies to doctors.
Information: https://www.propublica.org/series/dollars-for-docs
Search tool: https://projects.propublica.org/docdollars/
Data request: https://projects.propublica.org/data-store/sets/health-d4d-national-2

DocGraph Physician interaction network obtained through a freedom of information act request. Covers nearly 1 million entities.
Main page: http://www.docgraph.com
Information: http://thehealthcareblog.com/blog/2012/11/05/tracking-the-social-doctor-opening-up-physician-referral-data-and-much-more/
Data: http://linea.docgraph.org

---

5. UCI DATASETS

Liver Disorders Data Set

Data on 345 patients with and without liver disease. Features are 5 blood biomarkers thought to be involved with liver disease.
Data: https://archive.ics.uci.edu/ml/datasets/Liver+Disorders

Thyroid Disease Data Set

Data: https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

Breast Cancer Data Set

Data: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Heart Disease Data Set

Data: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Lymphography Data Set

Data: https://archive.ics.uci.edu/ml/datasets/Lymphography

Parkinsons Data Set

Data: https://archive.ics.uci.edu/ml/datasets/parkinsons

Parkinsons Telemonitoring Data Set

Data: https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set

Data: https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings

Parkinson's Disease Classification Data Set

Data: https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification

Primary Tumor Dataset Data: https://archive.ics.uci.edu/ml/datasets/primary+tumor

6. BIOMEDICAL LITERATURE

PMC Open Access Subset

Collection of all the full-text, open access articles in Pubmed central.
Information: http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
Archived files: http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#Data_Mining

PubMed 200k RCT

Collection of pubmed abstracts from randomized control trials (RCTs). Annotations for each sentence in the abstract are available.

Paper: https://arxiv.org/abs/1710.06071

Data: https://github.com/Franck-Dernoncourt/pubmed-rct

Web API of PubMed Articles

NLM also provided Web API for accessing biomedical literatures in PubMed.

Instructions for getting PubMed articles: https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PubMed/ (not full text, just title, abstract, etc.)

For articles in PubMed Central, instructions for getting the whole articles: https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC/

EBM NLP

Collection of pubmed abstracts from randomized control trials (RCTs). Annotation of Population, Intervention, and Outcomes (PICO elements) are available.

Paper: https://arxiv.org/abs/1806.04185

Data: https://ebm-nlp.herokuapp.com/annotations

Website: https://ebm-nlp.herokuapp.com/index

Evidence Inference

A dataset for inferring the results of randomized control trials (RCTs). A collection of pubmed RCTs from the open access subset. Annotations of (intervention, comparison intervention, outcome, significance finding, evidence span) are available.

Paper: https://arxiv.org/abs/1904.01606

Data: https://github.com/jayded/evidence-inference/tree/master/annotations

Website: http://evidence-inference.ebm-nlp.com/

PubMedQA

A dataset for biomedical research question answering. The task is to use yes/no/maybe to answer naturally occuring questions in PubMed titles.

Paper: https://arxiv.org/abs/1909.06146

Data: https://github.com/pubmedqa/pubmedqa

Website: https://pubmedqa.github.io/

6. TREC PRECISION MEDICINE / CLINICAL DECISION SUPPORT TRACK

Text REtrieval Conference (TREC) is running a track on Precision Medicine / Clinical Decision Support from 2014.

2014 Clinical Decision Support Track
Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.
Information and Data: http://www.trec-cds.org/2014.html

2015 Clinical Decision Support Track
Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.
Information and Data: http://www.trec-cds.org/2015.html

2016 Clinical Decision Support Track
Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records. Actual electronic health record (EHR) patient records are be used instead of synthetic cases.
Information and Data: http://www.trec-cds.org/2016.html

2017 Clinical Decision Support Track
Focus: Retrieve useful precision medicine-related information to clinicians treating cancer patients.
Information and Data: http://www.trec-cds.org/2017.html

7. MEDICAL SPEECH DATA

The TORGO Database: Acoustic and articulatory speech from speakers with dysarthria
The TORGO database of dysarthric articulation consists of aligned acoustics and measured 3D articulatory features from speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), which are two of the most prevalent causes of speech disability (Kent and Rosen, 2004), and matched controls. This database, called TORGO, is the result of a collaboration between the departments of Computer Science and Speech-Language Pathology at the University of Toronto and the Holland-Bloorview Kids Rehab hospital in Toronto.

Information and data: http://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

Paper: link

NKI-CCRT Corpus: Speech Intelligibility Before and After Advanced Head and Neck Cancer Treated with Concomitant Chemoradiotherapy.

NKI-CCRT corpus with individual listener judgements on the intelligibility of recordings of 55 speakers treated for cancer of the head and neck will be made available for restricted scientific use. The corpus contains recordings and perceptual evaluations of speech intelligibility over three evaluation moments: before treatment and after treatment (10-weeks and 12-months). Treatment was by means of chemoradiotherapy (CCRT).

Paper: http://lrec.elra.info/proceedings/lrec2012/pdf/230_Paper.pdf

Access: Contact authors.

Atypical Affect Interspeech Sub-Challenge

Björn Schuller, Simone Hantke, and colleagues are providing the EMOTASS Corpus. This unique corpus is the first to give access to recordings of affective speech from disabled individuals encompassing a broader variety of mental, neurological, and physical disabilities. It comprises recordings of 15 disabled adult individuals (ages range from 19 to 58 years with a mean age of 31.6 years). The task will be classification of five emotions from their speech facing atypical display. Recordings were made in their everyday working environment. Overall, around 11k utterances and around nine hours of speech are included.

Paper: http://emotion-research.net/sigs/speech-sig/is2018_compare.pdf

Link: http://emotion-research.net/sigs/speech-sig/is18-compare.

Autism Sub-Challenge

The Autism Sub-Challenge is based upon the "Child Pathological Speech Database" (CPSD) . It provides speech as recorded in two university departments of child and adolescent psychiatry, located in Paris, France (Universite Pierre et Marie Curie/Pitie Salpetiere Hospital and Universite Rene Descartes/Necker Hospital). The dataset used in the Sub-Challenge contains 2.5 k instances of speech recordings from 99 children aged 6 to 18

Paper: http://emotion-research.net/sigs/speech-sig/is2013_compare.pdf

Link: http://emotion-research.net/sigs/speech-sig/is13-compare.