Foundations for Integrating Data on Human Subjects without Compromising Individual and Institutional Privacy

Lucila Ohno-Machado, MD, PhD
University of California San Diego
Personalizing Prevention and Care

Gender: Male
Age: 20
Smoker: Yes
Total Cholesterol: 320 mg/dl
HDL Cholesterol: 20 mg/dl
Systolic Blood Pressure: 200 mm Hg
Medication Treatment for HBP: Yes
Variant: XYZ
Exposures: QWE
find all data about low grade astrocytomas with IDH mutations
A Closer Look at IDH Mutations | Astrocytoma Options

If you have a diagnosis of a grade II or III glioma (which includes astrocytomas, oligodendrogliomas and mixed oligoastrocytomas), I highly recommend asking ...
You've visited this page 2 times. Last visit: 3/7/17

Genetic Overview of Astrocytomas (WHO grades II and III ...

The following discussion applies mainly to astrocytomas of WHO grades II and III .... When “type 2” gliomas, that is, IDH-mutant astrocytomas without 1p/19q ...

IDH1 and IDH2 Mutations in Gliomas

by A Cohen - 2013 - Cited by 74 - Related articles
Mutations in isocitrate dehydrogenase (IDH) 1 and 2, originally discovered in .... These gliomas start as grade 2 astrocytomas, have the G-CIMP phenotype, and ...

IDH1 mutant malignant astrocytomas are more amenable to surgical ...

by J Beiko - 2014 - Cited by 80 - Related articles
Dec 4, 2013 - Patients with IDH1 mutant astrocytomas have a better overall prognosis compared to wild-type IDH astrocytomas, even after controlling for ...
Engaging The Community Toward a Data Discovery Index (DataMed v3.0)

Search for data through bioCADDIE

- Search for data set
- Search for repository

Statistics

- 74 REPOSITORIES
- 15 DATA TYPES
- 2,336,403 DATASETS
- 4 PILOT PROJECTS

Top 8 Repositories

- CCDC: 656,223 datasets
- Swiss-Prot: 438,182 datasets
- BioProject: 226,434 datasets
- Figshare: 192,552 datasets
- PDB: 131,204 datasets
- Dryad: 100,545 datasets
- OmicsDI: 78,201 datasets
- ArrayExpress: 69,996 datasets

New Features

- July 31, 2017. v3.0
  - API for DataMed
  - Autocompleting of words
  - Reporting of broken links
  - Context for search terms
  - Increase in number of repositories that are indexed

- Feb 28, 2017. v2.0
  - Increase coverage to more repositories
  - Duplicate datasets display feature
  - Usability enhancements based on user feedback and user interviews
  - User-reported issues resolved

Pilot Projects

- GWAS Finder
  - Search literatures for "Genome-Wide Association Studies". Read More

- iSEE-DELVE
  - Search Visualization project for Big Data. Read More

- DataRank
  - Find most suitable datasets for you. Read More

- Data Citation Discovery
  - Citation and Data Access Metrics Development applied to RCSB Protein Data Bank. Read More
Displaying 20 of 296 results for "astrocytoma grade 2 idh"

- **Genome-Wide Profiling of Astrocytic Gliomas**
  - **ID:** PRJNA119515
  - **Keywords:** Other
  - **Access Type:** download

- **Expression profiling of lower-grade diffuse astrocytic glioma**
  - **ID:** PRJNA150783
  - **Keywords:** Transcriptome or Gene expression
  - **Access Type:** download

- **Low grade gliomas subtype analysis**
  - **ID:** PRJEB9786
  - **Keywords:** Transcriptome or Gene expression
  - **Access Type:** download

- **Low grade gliomas subtype analysis**
  - **ID:** E-MTAB-3708
  - **Description:** Low grade gliomas (LGG; WHO grade 2 astrocytomas, oligodendrogliomas, and oligoastrocytomas)

- **Expression profiling of lower-grade diffuse astrocytic glioma**
  - **ID:** E-GEOD-35158
  - **Description:** Diffuse gliomas represent the most prevalent class of primary brain tumor. Despite significant recent advances in the understanding of gliomas, the genetic basis for differences between histological subtypes is incompletely understood. Here we present a novel, comprehensive transcriptomic analysis of 24 primary diffuse gliomas spanning the WHO grade 2-4 spectrum. We find that normal-appearing brain tissue surrounding glioma can be classified into three groups: tumors exhibiting astrocytoma or oligodendroglioma-like gene expression, tumors with features of intermediate differentiation, and pure dedifferentiated gliomas. A multiclass support vector machine analysis of these gene expression signatures correctly classified 96% of the gliomas and separated the three categories in a hierarchical manner.
Difficulties indexing study data

- Consents may not be allowed use by researchers outside a specific area (e.g., cancer)
- No participant unique identifier to connect data in different databases
- Data need to be indexable at the observation level, not only database level
  - Meta data for a database: title, author, abstract, year, etc.
  - Variables: demographics, diagnoses, geolocation, personal monitor data, etc.
find all data about low grade astrocytomas with IDH mutations
California Data for Site 1

Site 1

OMOP CDM Data

EHR System

Data

Site 2…N

OMOP CDM Data

EHR System

Data

Data aggregation

Data transfer > quarterly

Data & Research Support Center

OMOP CDM Data

Patient Reported

Wearables

NIH OT2OD024611
Clinical Studies vs. Electronic Health Records

• Informed consent is explicit in clinical and cohort studies
  • mental health, HIV status, substance abuse, sexually transmitted diseases may need to be itemized in the consent form
  • HIPAA authorization for EHR release
• Allowed usage is specified in most clinical studies
• Data are destroyed at completion of many studies

• EHRs that have undergone HIPAA “de-identification” do not require consent
Clinical Data Research Network

patient-centered Scalable National Network for Effectiveness Research

32 million people
14 health systems
How this works

• Federated data for distributed analytics

• Common data model

• Minimal computational infrastructure

• Compatible institutional policies

• Agreement on rules of engagement

• Shared ethics principles

• Research studies conducted on portions of the data
Big Data for Predictive Models (A.I.)

More genetic, sensor information now in EHRs

• ‘De-identified’ EHR data can be shared without patients’ extra consent
• Re-identification success from ‘de-identified’ data depends on many factors:
  • What is disclosed, to whom
  • Who an attacker is interested in identifying and how much she is willing to spend
HIPAA ‘De-identified’ data
- removal of 18 identifiers, such as dates, biometrics, names, etc.
- expert certification of low risk of re-identification
- ‘Limited’ data sets have ‘de-identified data’ plus dates
Genomes are Biometrics

- Biometrics are Protected Health Information (PHI)
- PHI requires HIPAA compliance
- Genomes should be treated as HIPAA identifiers

Are sensor data biometrics?
<table>
<thead>
<tr>
<th>Threat</th>
<th>System/Device</th>
<th>Risk</th>
<th>Defense</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD attack</td>
<td>Pacemaker, implantable cardioverter defibrillator, insulin pump, glucose sensor, intrathecal drug delivery, heart rate monitor, etc.</td>
<td>shorter effective lifetime</td>
<td>access control, shorter range communication</td>
</tr>
<tr>
<td>Replay attack</td>
<td></td>
<td>malfunction, irregular actions</td>
<td>time-stamping</td>
</tr>
<tr>
<td>DMB attack</td>
<td>Smartphones and their external mHealth devices</td>
<td>information leak, information injection</td>
<td>App-to-device level authentication</td>
</tr>
</tbody>
</table>

Ohno-Machado et al. Privacy, Security, and Learning in mobile health, commissioned article for the AAAS, 2014
Doing the Right Thing

Is it ethical to share?
People have not been explicitly asked and don’t know who is sharing what

Could people choose?
Is it practical?
What if massive number of people withdraw their data?

Is it ethical not to share?
New discoveries and acceleration of science depend on sharing
Personal monitors
no regulation of apps

EHRs
Disease, Family History, etc

Public ‘de-identified’
database of Stigmatizing
Condition X
Research database with “de-
identified” EHRs, genomes

Other databases
linking data for re-identification

My neighbor’s data
Does she have Disease X?
Privacy technology “solutions” (mitigation strategies)

**Institution/People-centric**
- Data broker (e.g., clinical data research networks)
- Patient-defined data sharing permissions (e.g., consent management)

**Data-centric**
- Add noise to data (e.g., differential privacy)
- Operate on encrypted data (e.g., homomorphic encryption)
- Multiparty computation (e.g., distributed analytics)
A National Center for Biomedical Computing focused on Data Sharing and Analysis
Institutional and Data-Centric Strategies

differential privacy
homomorphic encryption
secure multiparty computation

NIH U54HL1084600

Institutional and Data-Centric Strategies

- homomorphic encryption
- differential privacy
- secure multiparty computation
Distributed Computing

User requests data for Quality Improvement or Research

• Identity & Trust Management
• Policy enforcement

Diverse Healthcare Entities in 3 different states (federal, state, private)

Distributed Regression Model

First derivative:

\[ l'_r(\beta) = \sum_{i=1}^{n} \left\{ s_{i,r} - d_i \frac{\sum_{j \in R(t_i)} z_{j,r} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} \exp^{\beta^T z_j}} \right\} \]

\[ l'_r(\beta) = \sum_{k=1}^{m} \sum_{i=1}^{n} \left[ \sum_{z_j \in \text{site}_k} z_j - \sum_{i=1}^{n} d_i \frac{\sum_{k=1}^{m} \sum_{j \in R(t_i)} I(z_j \in \text{site}_k) z_{j,r} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} I(z_j \in \text{site}_k) \exp^{\beta^T z_j}} \right] \]

Hessian matrix:

\[ l''_{r,q}(\beta) = -\sum_{i=1}^{n} d_i \left\{ \frac{\sum_{j \in R(t_i)} z_{j,r} z_{j,q} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} \exp^{\beta^T z_j}} - \frac{\sum_{j \in R(t_i)} z_{j,r} \exp^{\beta^T z_j}}{\sum_{j \in R(t_i)} \exp^{\beta^T z_j}} \right\} \]

\[ l''_{r,q}(\beta) = -\sum_{i=1}^{n} d_i \left\{ \frac{\sum_{k=1}^{m} \sum_{j \in R(t_i)} I(z_j \in \text{site}_k) z_{j,r} z_{j,q} \exp^{\beta^T z_j}}{\sum_{k=1}^{m} \sum_{j \in R(t_i)} I(z_j \in \text{site}_k) \exp^{\beta^T z_j}} - \frac{\sum_{k=1}^{m} \sum_{j \in R(t_i)} I(z_j \in \text{site}_k) z_{j,q} \exp^{\beta^T z_j}}{\sum_{k=1}^{m} \sum_{j \in R(t_i)} I(z_j \in \text{site}_k) \exp^{\beta^T z_j}} \right\} \]

Conclusion: no patient data needs to be sent from the sites, only aggregates
Distributed Analytics across Horizontal and Vertical Partitions

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Sensor data</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>45</td>
<td>@$#(*$</td>
</tr>
<tr>
<td>A2</td>
<td>32</td>
<td>$&amp;*$(@</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient</th>
<th>Age</th>
<th>Sensor data</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>48</td>
<td>&amp;<em>(</em>)#**#</td>
</tr>
<tr>
<td>B2</td>
<td>72</td>
<td>#@$&amp;*(@</td>
</tr>
</tbody>
</table>

Li Y, et al. VERTIcal Grid IOgistic regression (VERTIGO)

*J Am Med Inf Assoc.* 2015
Big Data to Knowledge (BD2K) program
Phase II

1: Access Control
- User-Access Authentication
- Data-Access Authentication

2: Activity Ledger
- Critical User Activities
- Distributed Ledger

3: Risk Analysis
- Information
- Filtering

Analyze data to identify ancestry-specific variants associated with cancer

Cloud Service Provider:
- Amazon Web Services
- Azure
- Google Cloud Platform

IAM

Execution
Record
Blockchain
Privacy technology “solutions” (mitigation strategies)

**Institution/People-centric**
- Data broker (e.g., clinical data research networks)
- Patient-defined data sharing permissions (e.g., consent management)

**Data-centric**
- Add noise to data (e.g., differential privacy)
- Operate on encrypted data (e.g., homomorphic encryption)
- Multiparty computation (e.g., distributed analytics)
People-Centered Strategies

Patient Interface

Consent Management System
Do I wish to disclose data $D$ to $U$?

Sharing Look-up
I can check that $U$ looked at my data $D$

Healthcare Institutions

User $U$ requests Data $D$ on individual $I$

- Trusted broker
  - Data use agreements
  - Study registry

NIH R01HG008802
Informed CONsent for Clinical data Use in Research
# My Sharing Choices

<table>
<thead>
<tr>
<th>What clinical data am I sharing?</th>
<th>Who can access the clinical data I share?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>UCSD and VA Hospital</td>
</tr>
<tr>
<td>Age</td>
<td>✓</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>✓</td>
</tr>
<tr>
<td>Gender</td>
<td>✓</td>
</tr>
<tr>
<td>Race</td>
<td>✓</td>
</tr>
<tr>
<td>Socioeconomic Information</td>
<td>✓</td>
</tr>
<tr>
<td>Education Level</td>
<td>✓</td>
</tr>
<tr>
<td>Insurance Status</td>
<td>✓</td>
</tr>
<tr>
<td>Marital Status</td>
<td>✓</td>
</tr>
<tr>
<td>Occupation</td>
<td>✓</td>
</tr>
<tr>
<td>Income</td>
<td>✓</td>
</tr>
<tr>
<td>Sexuality</td>
<td>✓</td>
</tr>
<tr>
<td>Past Pregnancy</td>
<td>✓</td>
</tr>
</tbody>
</table>

Courtesy of H Kim
<table>
<thead>
<tr>
<th>Demographics</th>
<th>Current or Previous Disease or Condition</th>
<th>Therapy or Treatment Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Substance abuse related Disease or Condition</td>
<td>Medications</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Mental Health Disease or Condition</td>
<td>Social History</td>
</tr>
<tr>
<td>Gender</td>
<td>Sexual or Reproductive Disease or Condition</td>
<td>Alcohol Consumption Status</td>
</tr>
<tr>
<td>Race</td>
<td>Other</td>
<td>Recreational Drug Use</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smoking Status</td>
</tr>
<tr>
<td>Socioeconomic Information</td>
<td>Family's Current or Previous Disease or Condition</td>
<td>Health Care Encounter</td>
</tr>
<tr>
<td>Education Level</td>
<td>Substance abuse related Disease or Condition</td>
<td>Location of the hospital or clinic</td>
</tr>
<tr>
<td>Insurance Status</td>
<td>Mental Health Disease or Condition</td>
<td>Physician's Name</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Sexual or Reproductive Disease or Condition</td>
<td>Speciality of the hospital or clinic</td>
</tr>
<tr>
<td>Occupation</td>
<td>Other</td>
<td>Visit Dates</td>
</tr>
<tr>
<td>Income</td>
<td>Laboratory and Test Results</td>
<td>Charges and billing related to encounters</td>
</tr>
<tr>
<td>Sexuality</td>
<td>Genetic Test</td>
<td>Tissue and Blood Sample Usage</td>
</tr>
<tr>
<td>Past Pregnancy</td>
<td>Sexually Transmitted Disease Test</td>
<td>Tissue Sample</td>
</tr>
<tr>
<td>Anthropometrics</td>
<td>Test on Drug Screening</td>
<td>Blood Sample</td>
</tr>
<tr>
<td>Vital Signs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+ Sharing for stem cell research
iCONCUR: informed consent for clinical data and bio-sample use for research

Hyeoneui Kim,1 Elizabeth Bell,1 Jihoon Kim,1 Amy Sitapati,1,2 Joe Ramsdell,2 Claudiu Farcas,1 Dexter Friedman,3 Stephanie Feudjo Feupe,1 Lucila Ohno-Machado,1,4

ABSTRACT

Mean willingness-to-share score by individual data item

(HIV: HIV clinic, IM: IM clinic, Hx: History, FHx: Family History, Dx: Diagnosis, STD: Sexually Transmitted Disease, MD: Medical Doctor)
(a) Does having these choices make you feel differently about sharing your medical data?

- More willing to share: 59%
- No change: 35%
- Less willing to share: 5%
- Other: 1%

(b) How does your willingness to share your medical data for healthcare compare to your willingness to share for research?

- Don't know: 9%
- More willing for healthcare: 5%
- Equally: 7%
- More willing for research: 79%

(c) Would you feel more comfortable sharing your medical information if you know who is using it for research?

- More comfortable: 20%
- Indifferent: 4%
- Less comfortable: 3%
- Other: 73%

(d) If you could receive an update each time someone uses your medical data for research, would you like to be notified?

- Yes, each time: 24%
- Indifferent: 1%
- Don't need to know: 14%
- Quarterly or yearly: 61%
Big Data are a Big Deal

- App data: no regulation
- Databases: can be purchased
- Social network: can be used
- EHRs: can be linked

Asking people
Funding Sources

NIH R01HG008802

NIH U24AI117966

NIH T15LM011271

NIH U54HL1084600

Patient-Centered Outcomes Research Institute
CDRN-1306-04819

NIH OT2OD024611