Protecting and removing identifiers in human subject data

Dave Fearon, Sr. Data Management Consultant
JHU Data Services

GO TO dataservices.library.jhu.edu

EMAIL dataservices@jhu.edu

SHARE AT archive.data.jhu.edu

JOHNS HOPKINS LIBRARIES
Data Services

JHU DATA SERVICES

Protecting and removing identifiers in human subject data

Dave Fearon, Sr. Data Management Consultant
JHU Data Services

GO TO dataservices.library.jhu.edu

EMAIL dataservices@jhu.edu

SHARE AT archive.data.jhu.edu

JOHNS HOPKINS LIBRARIES
Data Services

JHU DATA SERVICES

JHU DATA SERVICES

HELPING YOU NAVIGATE DATA

WE HELP FACULTY, RESEARCHERS AND STUDENTS

FIND USE MANAGE VISUALIZE SHARE

FIND OUT MORE

GO TO dataservices.library.jhu.edu
EMAIL dataservices@jhu.edu
SHARE AT archive.data.jhu.edu

JOHNS HOPKINS LIBRARIES | Data Services

Before we start, a bit about ZOOM

zoom Video Conferencing

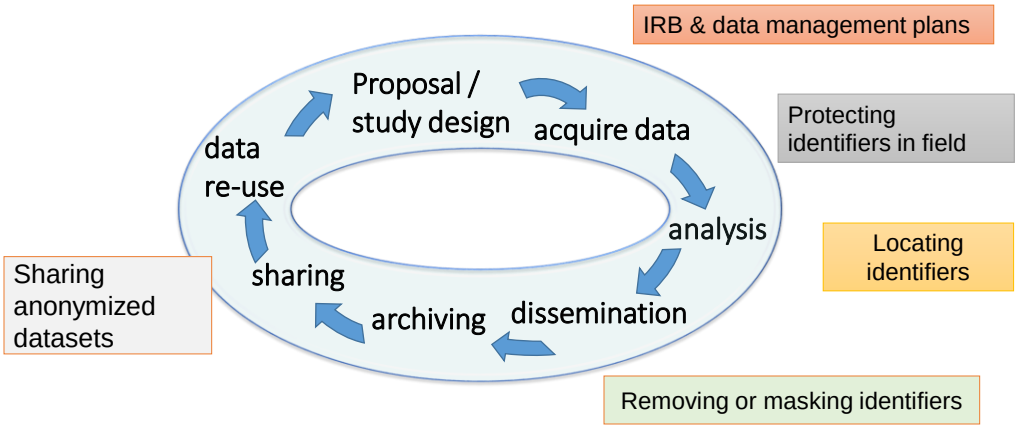
- Mute audio and video
- Ask a question
  - Use the public/private chat
  - Turn on your microphone (or press space bar) and speak up
- Interaction
  - Class polls
  - Feel free to leave your feedback in the chat

Learning Objectives

- How to locate and protect personal identifiers.
- When & how to prepare de-identified datasets for collaboration and sharing
- Terminology, intro to common techniques for research and collaboration
- Part 2: Advanced class with techniques and case examples

Consult with IRB and Data Trust (SOM) about compliance policies if planning to share de-identified data. (I am providing advice, they are the final authority)

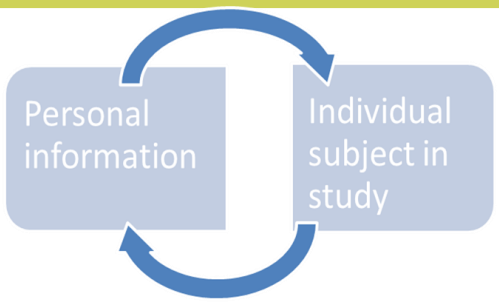
Protecting identifiers throughout the research process



What is identifiable data?

Personal and Health Identifiers

Information linked with individual participants they expect to remain private



PII: Personally identifying information

PHI: Protected Health Information

Personal and Health Identifiers

- **Names**: of subjects, related living people, employers
- **Identifying characteristics**: date of birth, images of subject, geographic locations
- **ID numbers** that permit links between individuals and their personal information
  - Social Security Number
  - Medical Record Number
  - Study ID Number you create for the project

Direct & Quasi-identifiers

Direct Identifiers: *uniquely private information*

Obvious	Less obvious
1. Names	2. Geographic division smaller than State (e.g. census tract)
3. Dates except year (e.g., birth date, date of research)	16. Biometric identifiers (fingerprint, voice recordings)
4-5. Phone, Fax No.	17. Face photos or comparable images
6. Email addresses	18. Any other unique identifying number, characteristic, or code
7. Social Security Numbers	
8-13. Medical & account numbers, licenses, vehicle/device numbers	
14-15 URLS, IP addresses	

Directly links variables to subjects, and to people or institutions associated with them.

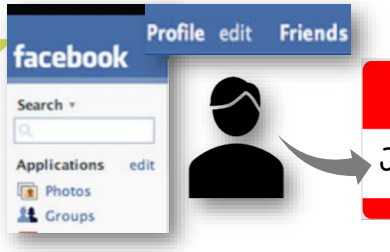
JOHNS HOPKINS LIBRARIES
Data Services

Direct & Quasi-identifiers

Direct Identifiers: *uniquely private information*

Some variables & data elements may be:

Indirect or **Quasi-identifiers:**
Event dates, locations, demographics, health measures…
Could link **some records** to **externally available information**



SubjID	Gend	HIVsts
2253	F	Pos
2254	M	Pos
2255	F	Neg

JOHNS HOPKINS LIBRARIES
Data Services

Quasi-identifiers can pose risk of linking to publicly available data

Example: database link re-identifies facility

Dataset	
Clinic Name	[deleted]
Clinic Revenue	$800-900K
Location	[deleted]

OK

Government Database	
Clinic name	Rosewood Rehab Center
Clinic Revenue	$824,209
Location	Baltimore, Maryland

Female
Pregnant
Veteran
Baltimore, MD

Unique "outlier" cases have higher probability of matching

Information not itself unique, but can be correlated with other information re-identify one or more participants in a study: e.g., dates, location, demographic info (race, ethnicity), or socioeconomic variables (occupation, salary)

Example 1: spot the identifiers: Medical records

JOHNS HOPKINS MEDICINE

Doe, Johnathan MR # JH65346 DOB: 07/15/1978

Doe, Johnathan
12/30/2013 8:00 AM Office Visit
MRN: JH653463

Description: **35 year old male**
Provider: **John Smith, MD**
Department: **Gss Som Orthopedics**

Diagnoses
Left ankle pain - Primary
719.47

Reason for Visit
Foot Pain
left heel

Vitals - Last Recorded

BP	Pulse	Temp(Src)	Ht	Wt	BMI
126/80	58	97.6 °F (36.4 °C) (Oral)	1.905 m (6' 3")	82.101 kg (181 lb)	22.62 kg/m2

Recent Review Flowsheet Data

Blood Pressure Percentiles by Age, Sex, and Stature	7/10/2013	7/16/2013	7/30/2013	12/30/2013
Blood Pressure	116/78	113/59	129/74	126/80

Treatment Plan
No notes of this type exist for this encounter.

Progress Notes
Ann Nonymous, MD : 12/30/2013 8:30 AM
Status: Sign at close encounter

He had a recent MRI (early December) at Union Memorial (Dr. Jones). While he does not have the imaging, he does have the report with him which does not indicate any gross abnormality of the Achilles tendon. He also tried hip/core strengthening exercise programs in the past (Union Memorial PT).

HPI: 35 y/o male with history of left heel pain since August of 2012. When the pain first started, he was trying to ramp up his running but found the pain increase at around the 10 to 12 mile mark when training for a marathon. He was diagnosed with an achilles tendonitis at that pont

He had one surgery on his left ankle (arthroscopic) in 2006 to shave off some bone while in graduate school at Penn State. At the time he was having trouble "cutting" while playing football and basketball. In 2007 he also had an avulsion fracture near to the sight of his current pain at the left heel.

Names (Patient & Dr.s)
Locations
ID numbers
Date of Birth & Age
Date of visit

Vitals?
Description identifiers?

Example 2: spot the identifiers, qualitative text

Oral History interviews from a study on Hurricane Katrina witnesses

From digital recordings originally posted online as "anonymous"

Subject: Okay I'm a New Orleans native. I've been in New Orleans basically my whole life.

Interviewer: Yes Ma'am

Subject: I'm a 55 year old black lady. And I was employed at the Dauphine Orleans hotel for 23 years my address was 415 Dauphine Orleans in the middle of the French Quarter. And when they had been announcing that there was gonna be a five category storm, cause that's there…

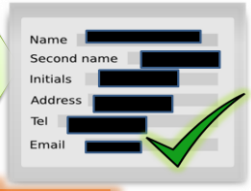
Interviewer: Yes Ma'am

Place of birth
Current residence
Age
Ethnicity
Place of employment
Address (workplace)

De-identifying data to reduce privacy disclosure risk

How to de-identify

Direct Identifiers: relatively simple to mask, most are not needed for analysis



Quasi-identifiers: more challenging to assess their risk and **anonymize** to decrease the probability of re-identification

Which variables are risky?

Date of Birth	Age	Date of onset
2/13/1928	92	~~2/13/2020~~

Removing interesting information?

JOHNS HOPKINS LIBRARIES
Data Services

De-identification varies in difficulty

How to de-identify

Some de-identification techniques are relatively simple

Birth Date	Age Range
~~2/13/1998~~	20-25

Advanced anonymization methods: unfamiliar & time consuming

K-anonymity risk probability calculations

$$\frac{1}{n}\sum_{j\in J} f_j \times I\left(\frac{1}{f_j} > \tau\right) \qquad \max_{j\in J}\left(\frac{1}{f_j}\right) = \frac{1}{\min_{j\in J}(f_j)}$$

JOHNS HOPKINS LIBRARIES
Data Services

The overall goal is to reduce the risk of privacy disclosure...

Reducing disclosure risk

Types of Disclosure Risk

Inappropriate Disclosure: attribution of information to a research subject or organization without their approval.

Three levels of disclosure risk

Identity disclosure	example
Subject can be directly identified, matched to a record	*MRN 213960.32 is Joe Biden*
Attribute disclosure	
Reveals information about subject, but not matching a specific record.	*Knowing person is in HIV study, that person may have HIV*
Inferential disclosure	
Released data makes it easier to determine a characteristic of a subject without linking to a specific record.	*Released variables commonly found on LinkedIn profiles*

HIPAA & privacy laws regulate Identity Disclosure, direct name matches

Attribute & Inferential may increase risk of directly matching records

Reducing disclosure risk

Case: Harvard Facebook matching

- 2006 study of 1700 Facebook profiles, many not publicly released, from "anonymous" university students
- 2008 - Dataset release from a Harvard repository, now restricted
- 2008 – U of WI privacy scholar Michael Zimmer cracked the location as: **Harvard's class of 2009**

Inferential disclosure	Linkable identifiers in *codebook*: size of class, major titles and housing systems unique to Harvard
Attribute disclosure	Posted (knowable) personal characteristics and preferences linkable to <u>some</u> but not necessarily <u>all</u> subjects.
Identity disclosure	Home state *outliers*: only 3 Utah students, directly identified with additional information

A few more disclosure protection efforts could have adequately protected this dataset (so it's not impossible!)

Reducing disclosure risk

What studies have disclosure risk?

Probably ready to use/share
Deceased subjects w/ no living relatives (Medical: 50+ years)
Public Use file – certified by IRBs, repository, gov. agency
Public opinion poll

Evaluate for Disclosure Risk	*examples*
Geographically specific	*Within a city or county*
Small samples	*organization-specific*
Purposive design	*longitudinal follow-up, snowball*
Matching external file	*city records database*
Sensitive content	*health or lifestyle risk factors*
Vulnerable subjects	*under age of majority (usually <16, 18)*
Detailed demographic, occupational, or biomedical variables (5+)	

Why de-identifying data?

Why de-identify data?

Consequences of disclosing personal & health identifiers

- Ethical first, protecting research participants
- Fines for institutions and researcher
  - (e.g. HIPAA regulated health identifier disclosure, $100-$50K fine per violation, up to $1.5 million and 10 years jail if for malicious intent.)
- Withdrawal of funding from institution, halting research
- Subjects can sue the institution
- Not good for one's career.
- However, demonstrating **due diligence** and **best practices** in protecting or removing identifiers can avoid or reduce penalties for a confidentiality breach (and reduce overall risk of breaches)

Why de-identify data?

Compliance with funder & publisher Data Sharing Policies

- Most US **federal funders**, some private funders, and many **publishers** have data sharing policies, or at least encourage sharing project data.
- NIH: Data sharing plans for all grants in 2023
- Funders **do not** require sharing data with human subject identifiers
- However, they may encourage efforts to remove identifiers for public access
- Some funders and grants require use of data repositories (e.g. **USAID, NIH genomics**)

Why de-identify data?

JHU Compliance and Data Stewardship when sharing data

JHU IRB

- IRB requires plans for protecting privacy for shared data.

SOM IRB

JOHNS HOPKINS SCHOOL of MEDICINE eIRB2

- SOM IRB Can require data sharing and de-identification plans, and preferred secure storage (SAFE Desktop)

Research Administration

- Data Use Agreements for external collaborations
- May require partial de-identification of shared data

Data Trust

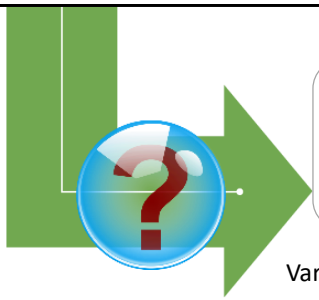
JHM Data Trust Research Subcouncil

http://intranet.insidehopkinsmedicine.org/data_trust

- Reviews requests for accessing data from JHM clinical, health plan, & business systems
- Approves **external data sharing** plans, including de-identification protocols
  - Protocols are reviewed by the CCDA

CORE FOR CLINICAL RESEARCH DATA ACQUISITION (CCDA)

How much de-identification is needed? (Depends on context of use)



Q: A researcher needs to share data with an external collaborator. She removed patient names and MRN, replaced with a code. Is it de-identified?
NO. (Have you used that term that way before?)

Varying definitions: Anonymized in Europe, De-identified in U.S. - could be treated as equivalent.

To de-identify is to protect against or minimize risk of re-identifying individuals from information

- Sometimes consists only of **masking**, where if key to the original data is known, information can be restored

To anonymize is to remove information by techniques and technical safeguards such that the data cannot be re-identified.

- a sub-category of de-identification applied especially to Quasi-identifiers

JOHNS HOPKINS LIBRARIES
Data Services

3 data use conditions, 3 de-identification levels

De-identify for context of use

	Data protection actions
Internal use by study team	**Remove direct identifiers not needed for analysis** Maintain secure internal access only
External Collaborators	**Remove direct identifiers and *most* quasi-identifiers** Maintain restricted access & secure data transfer
Public access sharing or database	**Remove direct identifiers and *all* quasi-identifiers** "Fully" anonymized, suitable for public access

JOHNS HOPKINS LIBRARIES
Data Services

De-identify for context of use

Internal use by study team

Data protection actions
Remove direct identifiers not needed for analysis
Maintain secure internal access only

- Remove identifiers or replace them with codes
  - E.g., Subject ID's, pseudonyms, category labels.
- Adds security to "working set" of data
- More secure storage and transfer within teams

Direct identifiers	
Nan **Subject ID**	Linked ID numbers: M **Event ID** numbers, licenses
Addresses	URLS, IP addresses
Phone, Cell	T **Clinic code** er locations
Email addresses	Photos/biometric IDs
Social Security Numbers	

Encouraged for **all research groups**

JOHNS HOPKINS LIBRARIES
Data Services

De-identify for context of use

Internal use by study team

Data protection actions
Remove direct identifiers not needed for analysis
Maintain secure internal access only

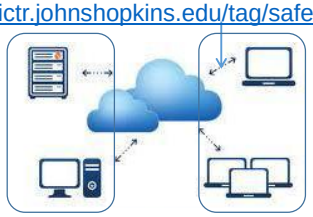
- Remove identifiers or replace them with codes
  - E.g., Subject ID's, pseudonyms, category labels.
- Adds security to "working set" of data
- More secure storage and transfer within teams
- Encouraged for all research groups

JOHNS HOPKINS LIBRARIES
Data Services

Preferred secure platform for PII/PHI

SAFE Desktop: Secure Analytic Framework

https://ictr.johnshopkins.edu/tag/safe-desktop/



De-identify for context of use

External Collaborators

Data protection actions
Remove direct identifiers and most quasi-identifiers
Maintain restricted access & secure data transfer

- Broaden specific values for quasi-identifiers
  - E.g., Numerical values to ranges: Age 52 → 50-55
- Share with Data Use Agreements with IRB-approved researchers
- Secure data transfer and storage required

- **Requires restricted access, secure data transfer**

JOHNS HOPKINS LIBRARIES Data Services

OneDrive SharePoint REDCap™

De-identify for context of use

External Collaborators or **Restricted Data Repository**

Data protection actions
Remove direct identifiers and most quasi-identifiers
Maintain restricted access & secure data transfer

- Restricted Data Repositories protect and manage access to deposited datasets
- Usually requires removal of most PII/PHI in data

Restricted Data Repositories:

Genomics:	dbGaP
Mental Health:	NIHM Data Archive
Social Sciences:	ICPSR
Public Health:	USAID DDL

Require partial de-identification

JOHNS HOPKINS LIBRARIES Data Services

De-identify for context of use

Public access sharing or database

Data protection actions
Remove direct identifiers and *all* quasi-identifiers
"Fully" anonymized, suitable for public access

- A "fully de-identified" or "anonymous" dataset, no longer contains "human subject" data.
- Apply advanced statistical de-identification techniques if needed.
- Often requires professional assistance. Rarely a do-it-yourself procedure for public release.

Can share directly or via "open access" data repositories

JHU Data Archive
http://archive.data.jhu.edu
Contact dataservices@jhu.edu to archive data.

Another context for de-identification levels

De-identify for context of use

HIPAA levels for data de-identification

Internal use by study team	**Limited Dataset (LDS)**
External Collaborators & restricted repositories	**"Safe Harbor" Datasets**
Public access sharing or database	**Expert/Statistical Determination data = "De-identified"**

HIPAA
Health Insurance Portability and Accountability Act

De-identify for context of use

Internal use by study team

Data protection actions
Remove direct identifiers not needed for analysis
Maintain secure restricted access

Data can include PII/PHI w/secure access, but when possible, aim toward:

SAFE Desktop

Limited Dataset (LDS)

- ⊘ Direct identifiers removed or masked
- ✓ Zip code-level location
- ✓ Full dates can be retained

2/13/1998

HIPAA Health Insurance Portability and Accountability Act

JOHNS HOPKINS LIBRARIES Data Services

For IRB/compliance forms, do not refer to Limited datasets as "de-identified"

De-identify for context of use

External Collaborators or **Restricted Data Repository**

Remove direct identifiers and most quasi-identifiers
Maintain secure restricted access

Limited Datasets share w/ external collaborators only under DUA's. Compliance offices prefer sharing: "Safe Harbor" Datasets

- ⊘ Direct & quasi-Identifiers that "knowingly" re-identify
- ✓ 3-digit Zip code truncation
- ✓ Year only dates, ages above 90

HIPAA Health Insurance Portability and Accountability Act

Requires restricted access, secure data transfer

OneDrive REDCap

Birth Date	Age	
~~2/13/1927~~	~~94~~	90+ ✓

De-identify for context of use

External Collaborators or **Restricted Data Repository**

Remove direct identifiers and most quasi-identifiers
Maintain secure restricted access

Limited Datasets share w/ external collaborators only under DUA's. Compliance offices prefer sharing: "Safe Harbor" Datasets

For JHU SOM and HIPAA covered entities

- JHM data often requires **risk review** & approval by the **JHM Data Trust Council**
- Review of de-identification protocols for **Safe Harbor** datasets by the CCDA (Core for Clinical Research Data Acquisition)
- Submit collaboration details to IRB with planned de-identification level

De-identify for context of use

Public access sharing or database

Remove direct identifiers and all quasi-identifiers
"Fully" anonymized, suitable for public access

Expert/Statistical Determination = "De-identified"

- ✓ De-identification performed with appropriate knowledge, accepted statistical techniques
- ✓ Data assessed for remaining disclosure risk
- ✓ Documented methods and results of analysis

For **JHU Archive** and **public access repositories**: IRB may also require Data Trust & CCDA review for JHM patient-derived data

Know the de-identification level of your dataset

Research study members should ideally be able to:

- Understand which identifiers have high risk
  (e.g., variables matching external public information)
- Avoid collecting direct identifiers not required for research
- Create a "**working copy**" of data with unnecessary identifiers removed, ideally to "Limited Dataset" levels.
- Maintain **restricted access** to any data derived from PHI/PII

Utility Maintain analytic utility while protecting subjects

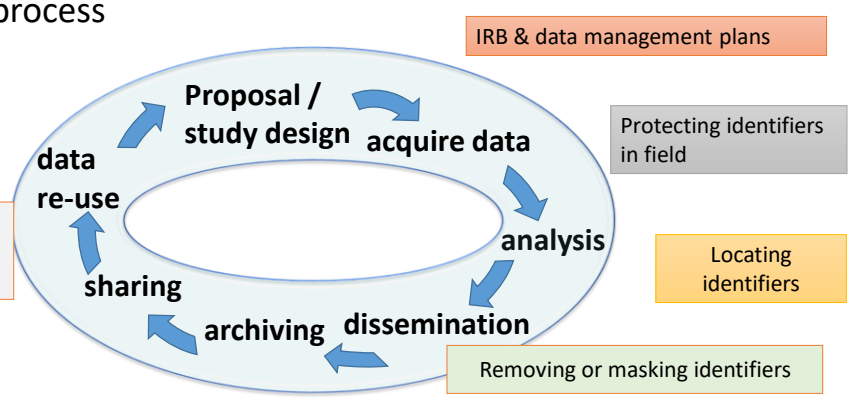
- Obliterating all risk may make data unusable for others
- Restricted access limited data sets are often the best choice

Preparing for disclosure protection throughout the research project

Preparing for disclosure protection throughout the research project

Removing identifiers can require significant time and effort, so

Plan ahead: prepare de-identified sets before the project is complete, at each stage of the research process



JOHNS HOPKINS LIBRARIES
Data Services

Plan ahead: disclosure protection and sharing in IRB Forms

There are two places in your IRB forms where data sharing plans should be addressed:

Research Plan

Consent Form

IRB Research Plan (eIRB)

- **What data will be shared?**
  - *How* - peer to peer, via a public repository
  - *When* - after a project, post-publication
  - *Types* - raw data, analyzed subset, transcripts, audio/video
- **How will you protect the rights and privacy of human subjects** both during and after the study?
- Who can use data and under what conditions
- **How long will data be retained?**
  - *How will files with identifiers be disposed of?*

JOHNS HOPKINS LIBRARIES
Data Services

SOM IRB: May require **Security Checklist** for certain data sharing or data re-use (eFormB) requests. And review by the **JHM Data Trust Research Subcouncil** of de-identification protocols for external collaborations

Data Trust

Consent Form Language

- In most cases, it is essential to get participant's consent to share data online in a repository, even for restricted repositories and for most de-identified datasets

Statement examples:

State where de-identified data will be shared:	Selected data from this study will be deposited into the Johns Hopkins Data Archive, a publicly-accessible database for research data. We will remove all identifiers linked to you before the data are deposited.
Inform them of small risk:	There is a very small chance that someone with access to the research data or results could identify you through external information sources. JHU researchers have policies and practices in place to minimize any risk of indirect disclosure of your personal information.

JOHNS HOPKINS LIBRARIES
Data Services

Work with your IRB team to develop a consent form for data sharing.

Consent Form Language with Opt-Out Clause

- Allowing participants to opt-out of sharing de-identified data:

☐ I do not wish to have anonymous transcripts shared online or through a data repository for further research or educational purposes, even though there is a low risk that I can be identified by the information released.

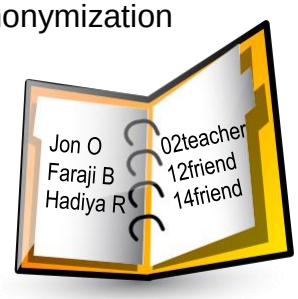
- Consider emphasizing the low risk of sharing de-identified data vs. the value of sharing

JOHNS HOPKINS LIBRARIES
Data Services

Participants can opt in on identifying less sensitive info
e.g., voice on audio recordings

Protect Identifiers During Data Collection

- **When collecting data in fieldwork or clinical settings**
  - If recruiting participants, keep their contact info secure and separate from materials you bring to the field
  - Prior to data collection, prepare an anonymization scheme and/or secure key code list
  - Document identifiers and their substitutions in a secure list or codebook…



Review the Codebook

- Review Codebook or Data Dictionary that defines variables.
- Note direct and especially quasi-identifiers to check for the data & protect if collected or shared

Variable	Codes	Label
P2cluster	ID	Cluster ID
P2district	Baltimore, DC, Fredrick, Towson	VDC code
P2surveydate	Survey date	Survey date
P2womana	age in years	Woman's age
P2womansch	Years of schooling	Completed years schooling
P2childbenefits	Benefit category	Child benefits

location

dates

Subject attributes

public database?

JOHNS HOPKINS LIBRARIES

Data Services

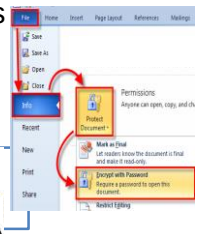
Protect Identifiers During Data Collection

Data security for workstations, laptops, and mobile devices

- Keep mobile devices in locked secure places
- In remote locations, backup files to **secure cloud storage** ASAP, however:

Encrypt identifier files before they are stored or transmitted:

- Add password protection to MS Office or other files
- **VeraCrypt** or other programs can encrypt folders
- Secure the encryption passwords, share with an approved collaborator/advisor.



Encrypted USB and hard drives: software or hardware based

Protect Identifiers During Data Collection

SAFE Desktop: Secure Analytic Framework Environment

- #1 for IT risk compliance. Software runs within virtual server. 100GB storage (increasable)
- Only JHED ID access.

https://ictr.johnshopkins.edu/tag/safe-desktop/

REDCap™

- HIPAA compliant for online surveys and data management
- Allows external collaborators. Paid levels get JHU staff support.

OneDrive

Microsoft Teams

- HIPAA compliant security IF shared access is managed properly
- Allows external collaboration, PHI storage not recommended.

Never Dropbox!

IT@JH NAS servers preferred over **Department-managed Servers for JHU Medical data or Sensitive Data**

Analysis phase: marking & changing identifiers

Optimal times for locating identifiers are shortly after data collection, or during data analysis.

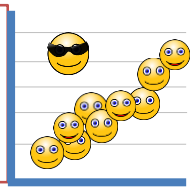
When you see an identifier, change it or tag it to change later

Untitled - SPSS Data Editor

File Edit View Data Transform Statistics Graphs Utilities Window Help

SubjID	Gend	Region	RegGen	HIVsts	De-ID
2253	F	21239	212	Pos	Y
2254	M	21238	211	Pos	Y
2255	F	21135	211	Neg	N
2256	M	06058	065	Neg	Y

*Watch out for **Outliers!*** Mark records with unique or uncommon combinations of variables. These can become **quasi-identifiers** if linked to outside info.



Goal is a **working version of data for analysis** with mostly de-identified variables either in use or marked for removal for shared versions

Preparing de-identified datasets for sharing

Preparing de-identified datasets for sharing

Before sharing data in any form, prepare a version that is de-identified sufficiently for how it will be accessed:

Internal use by study team — **Remove direct identifiers not needed for analysis**

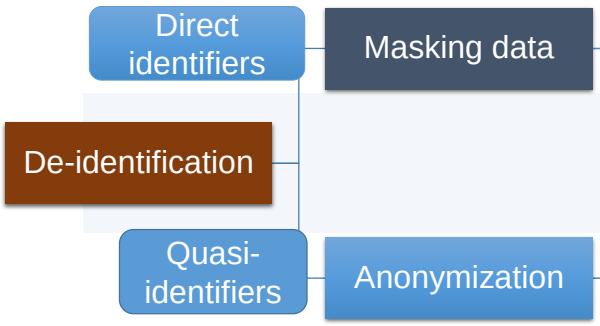
External Collaborators, restricted repositories — **Remove direct identifiers and *most* quasi-identifiers**

Public access sharing or open access data repository — **Remove direct identifiers and *all* quasi-identifiers**

Basic techniques are often sufficient

more advanced techniques (Part 2 webinar)

De-identification techniques



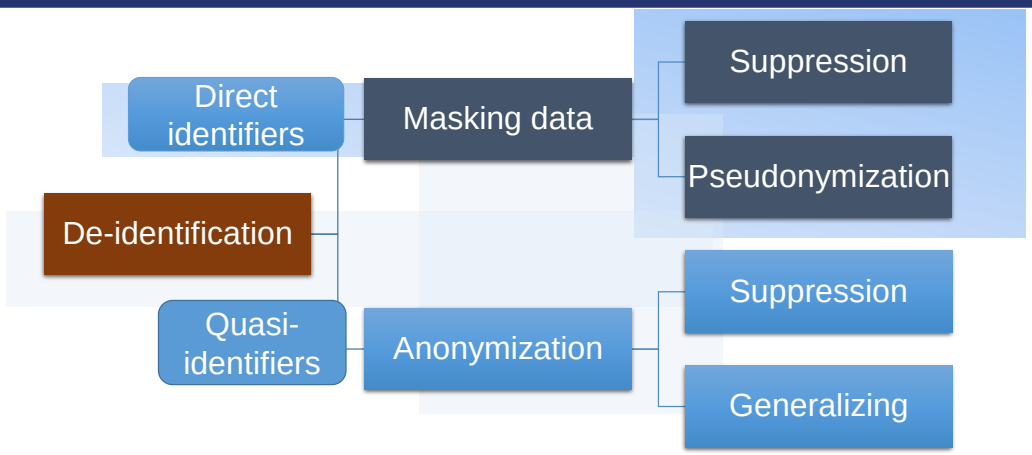
Masking: *removing or substituting direct identifiers*

- Often if the **key** to the original data is known, information can be restored

*To **anonymize**: removing information by techniques & technical safeguards to prevent re-identification.*

- applied especially to Quasi-identifiers

De-identification techniques



What type of identifier is rarely if ever needed for analytic purposes?

- Direct identifier
- Quasi identifier
- Sensitive identifier

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

JOHNS HOPKINS LIBRARIES

Data Services

De-identification steps:

Step 1) Review and Mask Direct Identifiers

Data masking (defined) techniques for reducing the risk of identifying a data subject to a negligible level for data with no utility for analysis in its original form.

Antenatal Card: [redacted] B2-2296 [redacted] Report Date: 2006-[redacted]

Patient Name	Patient Id	Age	Birth Date	Address	Husband	Clinic Name
[redacted] Patient2332	[redacted]	[redacted]	[redacted]	[redacted]	[redacted] married [redacted]	Inst_OC3

Lab Tests

Date Request	Lab Type	Date Results	Results	Site	Clinician
2006-[redacted]	Hb - 1st screen	2006-[redacted]	[redacted]	InstOC3	[redacted] Dr_16 [redacted]

Drug Interventions

Date Given	Folic	Iron
2006-[redacted]	X	X

- Primarily for **Direct Identifiers**, but are also used for removing Quasi-identifiers from analysis.

Direct identifiers
Names
Addresses
Phone, Cell
Email addresses
Government/National ID Numbers
Linked ID numbers: Medical & account numbers, licenses, etc.
URLS, IP addresses
Treatment provider locations
Photos/biometric IDs

JOHNS HOPKINS LIBRARIES

Data Services

Data masking techniques

Field suppression

Removing a value from the data set or replacing a value with a NULL.

MRN	Email	Age	Ethn.	Diagnosis Code	Type2-DB onset
293.3506		[redacted]	W	**0**	83
495.4649		45	W	**0**	[redacted]
384.5498		33	W	**0**	30
[redacted]					
399.5499		89	B	**0**	70

- The simplest data masking method
- Typically applied to field values, but cells or rows can be suppressed
- Effectively removes the whole variable that can potentially be used to re-identify records.

Data masking techniques

Pseudonymization

- Replaces with a unique value that is maintained consistently for each use
- Name pseudonyms are typical for qualitative text or transcripts
- Pseudonyms used as **key field identifiers** can maintain table relationships.

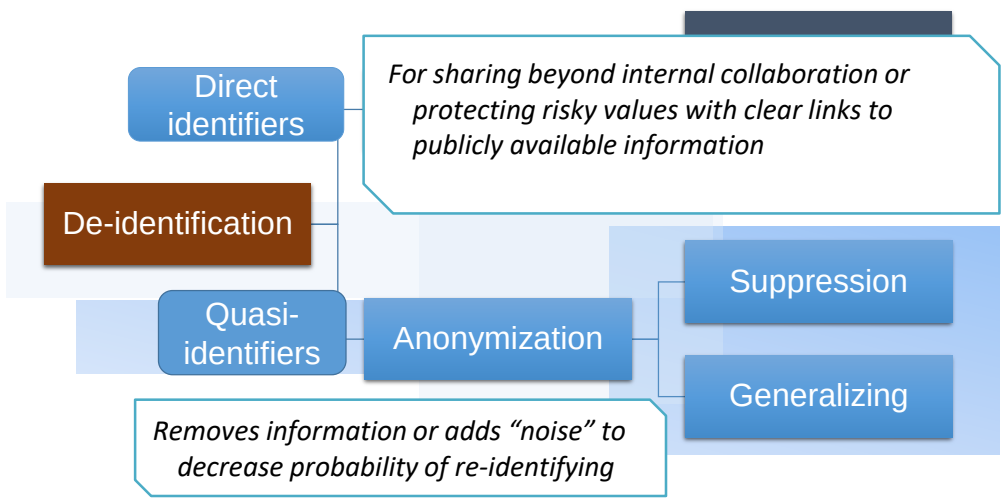
MRN	Last Name	CSN
17274.685	**Mathews**	**924**
16303.342	Jones	456
18048.841	McGaren	333

Visit	[table2]
MRN	16303.342
CSN VisitKey	456
Admit Date	20190323

Test	[table3]
CSN	456
BP	120/60

Replacement keys for Pseudonyms must be maintained securely, or names can be re-associated.

De-identification techniques



Step 2) Generalize specific dates

Generalization — techniques for reducing the precision of a value to a more general one (e.g. range vales, date shift, broaden categories.)

Date identifiers
Specific day
Date of Birth
Date of interview
Date of treatment
Dates that can be found in public records

Data Services

- Generalize dates to maintain analytic utility:
  - Full date → Month/year → Year
  - Broaden day/month to season
- Often necessary that changed dates maintain duration or subsequent events:
  - Change Date of Birth to Age or generalize to Age Range
  - Calculate duration from date sequences
  - **Date shifting**…

Birth Date	Age	Age Range
2 removed	21	20-25

Visit Dates	Days b/t visits
3/2/2017	**0**
4/9/2017	38
5/8/2017	29
7/2/2018	420

Generalize specific dates

Date Shifting

Fixed data shift

- Shift the entire dataset by 15 days
- Risks guessing the pattern

Date	+15Days	Randomized date
3/2/2018	**15**	3/17/2018
4/9/2017	**15**	4/24/2017
5/8/2018	**15**	5/23/2018
7/2/2018	**15**	7/17/2018
8/2/2018	**15**	8/17/2018

Randomize the date shift within some set range

- Smaller ranges may be sufficient for restricted access
- +/- 180 days for **Safe Harbor**

Date	+-15Days	Randomized date
3/2/2018	**11**	3/13/2018
4/9/2017	**-15**	3/25/2017
5/8/2018	**9**	5/17/2018
7/2/2018	**-3**	6/29/2018
8/2/2018	**12**	8/14/2018

Step 3) Remove or anonymize geographic variables

- Retain geography only to level required for analysis, within the risk threshold.

Geographic variables to remove or recode
Street Address
Census tracts
ZCTA (zip code tabulation areas)
County
Congressional Districts
Urban neighborhoods
Area populations < 100,000

Data Services

Restricted access among collaborators:

Remove direct addresses, generalize Zip code to 3 digits

Postal code:

21219

For full de-identification:

Areas ≥ State/province

Categorize regions: urban/rural

population blocks ≥ 100,000

Advanced: GIS polygon mapping

Generalization of other data types

Top and bottom coding:

- change extreme top & bottom of outlier variables:

Age	Actual Wealth	Top-coded Wealth
24	24,778	24,778
31	26,750	26,750
42	26,780	26,780
64	35,469	**30000+**
27	43,695	**30000+**

Collapse categories with low frequencies (low *p* value)

- make broader ranges by creating a broader coding scheme.

Mineral	#	Protein	#	Vitamin	#
Potassium	1	Kwashiokor	0	B2	3
Magnesium	2	Marasmus	3	B12	0
Calcium	5	Catabolysis	1	C	2
Zinc	0			A	0

Deficiencies	#
Mineral	8
Protein	4
Vitamin	9

Child health	#
Malnutrition	21

JOHNS HOPKINS LIBRARIES
Data Services

Step 4) Remove or anonymize **quasi-identifiers** that pose risk of link to external datasets

- Example: Geographic ID revealed by database link

Survey	Response
Methadone outpatient case#	< Name deleted >
County Clinic Revenue	**$800-900K**
Location	<deleted>

Matching value

Linked location

District	Database
District clinic facility name	Rosewood Rehab Center
County Clinic Revenue	**$824,209**
Location	Rosewood, MD

JOHNS HOPKINS LIBRARIES
Data Services

- Change: Recode revenue as range or grouped averages

Remove potential links of no analytic value

Weigh risks/utility of removing/recoding based on likelihood of external link

When to consider advanced de-identification techniques

- Applying the prior 4 steps is often sufficient disclosure protection for **restricted access data depositories**
- Consider more advanced de-identification techniques when:
  - preparing a **public access dataset**
  - de-identification removes too many key variables of interest
  - preparing complex quantitative datasets, especially for large samples with multiple variables
- May require a **professional statistician**, especially for public access datasets, and JHU Honest Broker for HIPAA-covered data

JOHNS HOPKINS LIBRARIES
Data Services

Broad-brush approach: never publicly share datasets requiring advanced statistical anonymization

Moderate approach: evaluate Quasi-ID risk for context of release with a basic risk threshold measure

Setting the risk threshold

K-Anonymity model (Latanya Sweeney)

K= (roughly) the set of records with matching values for all quasi-identifiers, to calculate the probability of re-identifying a particular record.

K=2 to K=9
Acceptable thresholds for **restricted access repositories**

Risk Threshold

0 K=2 — **SAFE desktop**

0.5 K=2 — **PMAP**

0.33 K=3 — **Redcap** collaboration

0.2 K=5

0.11 K=9 — **Vivli, NIMH Data Archive** restricted

0.09 K=11 — **JH Data Archive** Open Access

0.05 K=20

Public release: K=11 to K=20 matching records for combined Quasi-ID variables.[1]

Internal collaborations
Minimal de-identification
Higher data quality

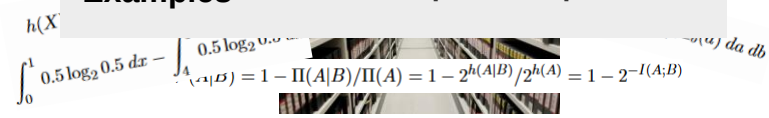
Higher risk context
Significant de-identification
Lower data detail

[1]Emam, Khaled El, and Luk Arbuckle. 2013. *Anonymizing Health Data*. O'Reilly Media, Inc.

Resources for Statistical Disclosure Protection

- Learning advanced statistical de-identification techniques requires some study:
  - Books and journal articles (at the library!)
  - **Very few** internet resources for learning advanced techniques

Part 2 webinar: De-identifying Human Subject Data: Techniques and Case Examples **Dec 3 @ 12:00 pm – 1:00 pm** **Online**



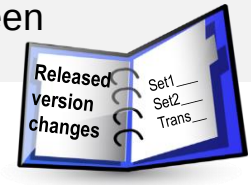
JOHNS HOPKINS LIBRARIES
Data Services

JHU Data Services dataservices@jhu.edu can help assess de-identification strategies for your project

Documenting & closing projects

Documentation: add a summary of release procedures

- **Always document shared datasets**: Important to summarize what has been changed from the original data
  - Necessary for replication of study
  - Helps account for variances



- To the extent possible; be careful not to compromise disclosure protection procedures
- For **IRB**: Indicate on the annual progress report that the dataset has been de-identified

Completing a project: What to do with identifiers? That all depends

If identifiers are needed for future study:
e.g., longitudinal or comparative
- Keep identifier set **secure**
  - *encrypted, secure password*
  - *off networks, single backup*
- Keep de-identified set separate.

Follow original IRB research plan for identifiers, or update any changes

If identifiers are not needed for future study:
e.g., sensitive, no follow-ups
- **IF** de-identified sets retain most of the utility for re-use & verification…
- destroy the identifier sets & their backups

Complete the documentation of both de-identified & identified datasets

Decide **who is responsible** for the datasets – long term!
- Primary and 2nd person to handoff responsibility if needed
  Who can you trust with the identifiers? You are the custodian.

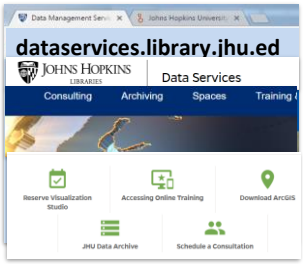
In closing, for human subject research:

Planning from the beginning of research is critical for
- Protecting identifiers
- Disclosure analysis & de-identifying
- Sharing publicly accessible data

Budgeting for de-identification
- Cost in time & labor
- Extra staffing required?
- Protecting restricted data

But your efforts can pay off with a dataset with high utility for your research community

JHU Data Services can provide consultative advice and point you to available resources



Disclaimer: We are providing advice; IRB & research compliance offices are the final authorities on this subject

Contact JHU Data Services

GO TO
dataservices.library.jhu.edu

EMAIL
dataservices@jhu.edu

SHARE DATA AT
archive.data.jhu.edu

De-identification resources

See website for upcoming webinars

Webinars: De-identifying Human Subject Data for Sharing
October 18 @ 12:00 pm – 1:00 pm Online

Part 2: De-identifying Human Subject Data: Techniques and Case Examples
Dec 3 @ 12:00 pm – 1:00 pm Online

Online guidelines and training modules: (see our website)

Consulting: Planning for data de-identification and risk screening

JHU Data Services: for managing, sharing, finding, visualizing data, and geospatial research support

FIND OUT **More**

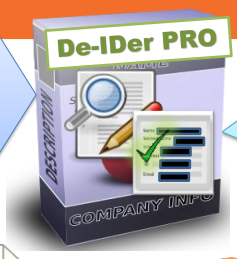
https://youtube/r4VqNMw6Q8s

Johns Hopkins Libraries | Data Services

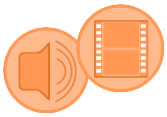
De-identification software

Can software de-identify for us?

Locating identifiers that risk disclosure



Altering the data to remove risk



- Relatively few applications, many open source, minimally supported, or enterprise-level ($$)
- Most require some expertise in disclosure protection methods to use correctly

De-identification software list

https://dataservices.library.jhu.edu/resources/

Applications to Assist in De-identification of Human Subjects Research Data

- Tools for De-identifying Unstructured Text
- Tools for De-identifying Data in Digital Images
- Tools for De-identifying Tabular or Otherwise Structured Data

JOHNS HOPKINS LIBRARIES

Data Services

Structured tabular data

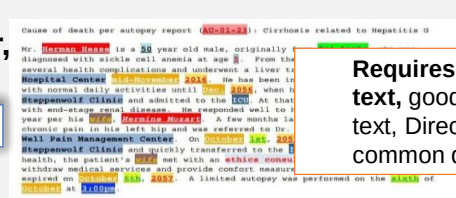
The sdcMicro package in R https://cran.r-project.org


```
# load protected data (as created in the
example # of \code{\link{protectTable}})
sp <- searchpaths() fn <-
paste(sp[grep("sdcTable", sp)],
"/data/protectedData.RData", sep="")
 protectedData <- get(load(fn))
```

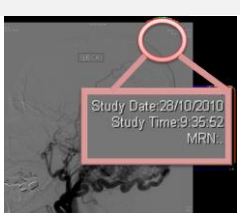
Might be too advanced for risk assessment workflow but option for R deposits.

Unstructured text: NLM-Scrubber, National Library Medicine

https://scrubber.nlm.nih.gov



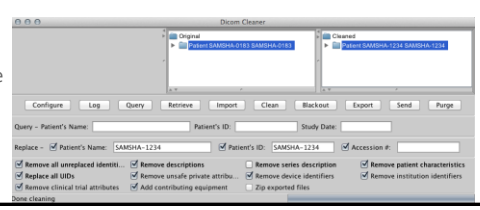
Requires setup to prep text, good for structured text, Direct identifiers, common quasi-identifiers



Medical imagery

•DICOMCleaner

- Software description: "DicomCleaner™ is a free open source tool with a user interface for importing, ""cleaning"" and saving sets of DICOM instances (files)"
- Intended purpose: Medical Images in DICOM format, such as radiology



De-identifying qualitative data

Removing identifiers in **qualitative text data**:

Challenge	Locating direct & quasi-identifiers in text
Solutions:	• Use software to make global changes to regular expressions (e.g. names)

- Make changes manually when ID's are encountered during analysis
- Remove or replace uniquely identifying words and phrases:

Substitute identifiers with variable code value
[Paraphrased text in brackets]
Mark deleted sections [description of event removed]

Subject: Okay I'm a **[Region1City]** native. I've been in **[Region1City]** basically my whole life.

I'm a **[50-70agerange]** year old black lady. And I was employed at the **[cityhotelname]** hotel for 23 years my address was **[gives address]** in the middle of the **[tourist area].**

[paraphrase: evacuated to brother's house…]

JOHNS HOPKINS LIBRARIES

Data Services

Masking audio & video recordings

Challenge:	Audio of voices is considered inherently identifiable. Photos and video difficult to mask.
Solutions:	• Seek **consent form** approval for restricted release of audio/video clips

- Off-the-shelf AV editing software can blur images and disguise vocal audio (reasonable workload for small batches of sample clips)

JOHNS HOPKINS LIBRARIES

Data Services