


GO TO
dataservices.library.jhu.edu

EMAIL
dataservices@jhu.edu

SHARE AT
archive.data.jhu.edu

Protecting and removing identifiers
in human subject data

Dave Fearon, Sr. Data Management Consultant
JHU Data Services




JHU DATA SERVICES

© 2022. For distribution to JHU Affiliates only


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

Do not distribute beyond JHU affiliates
without permission. © 2022

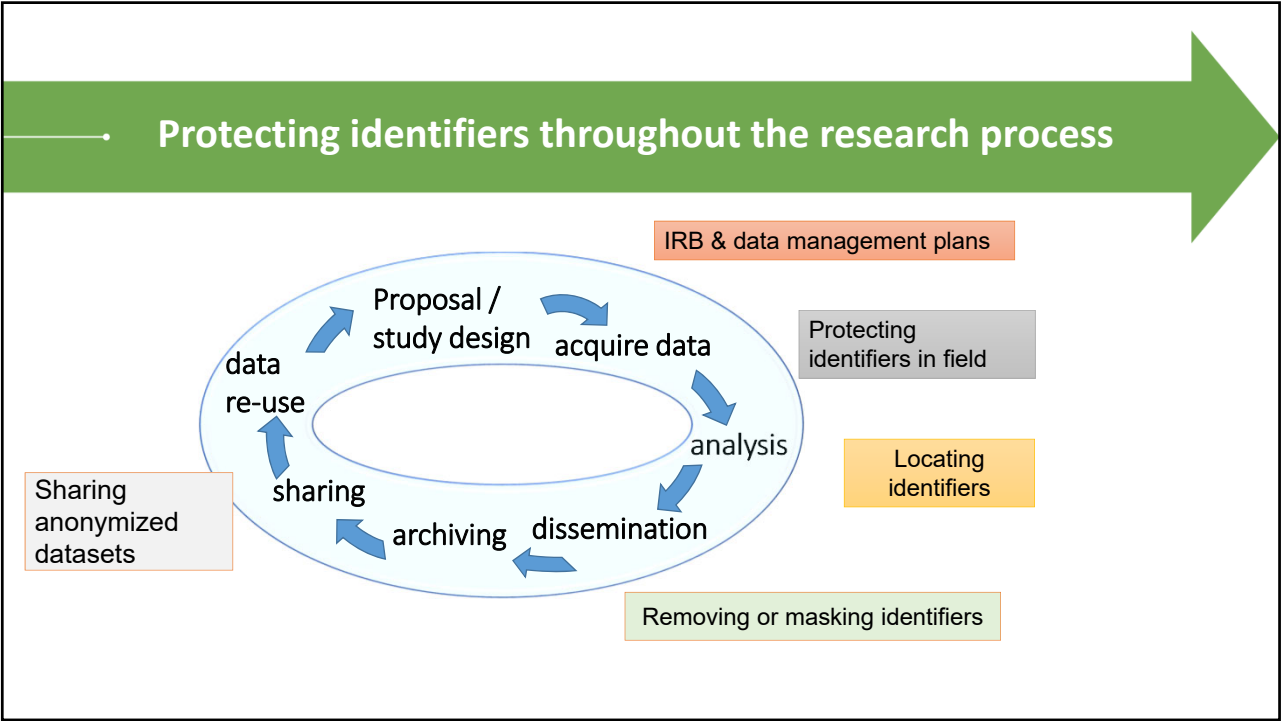
1

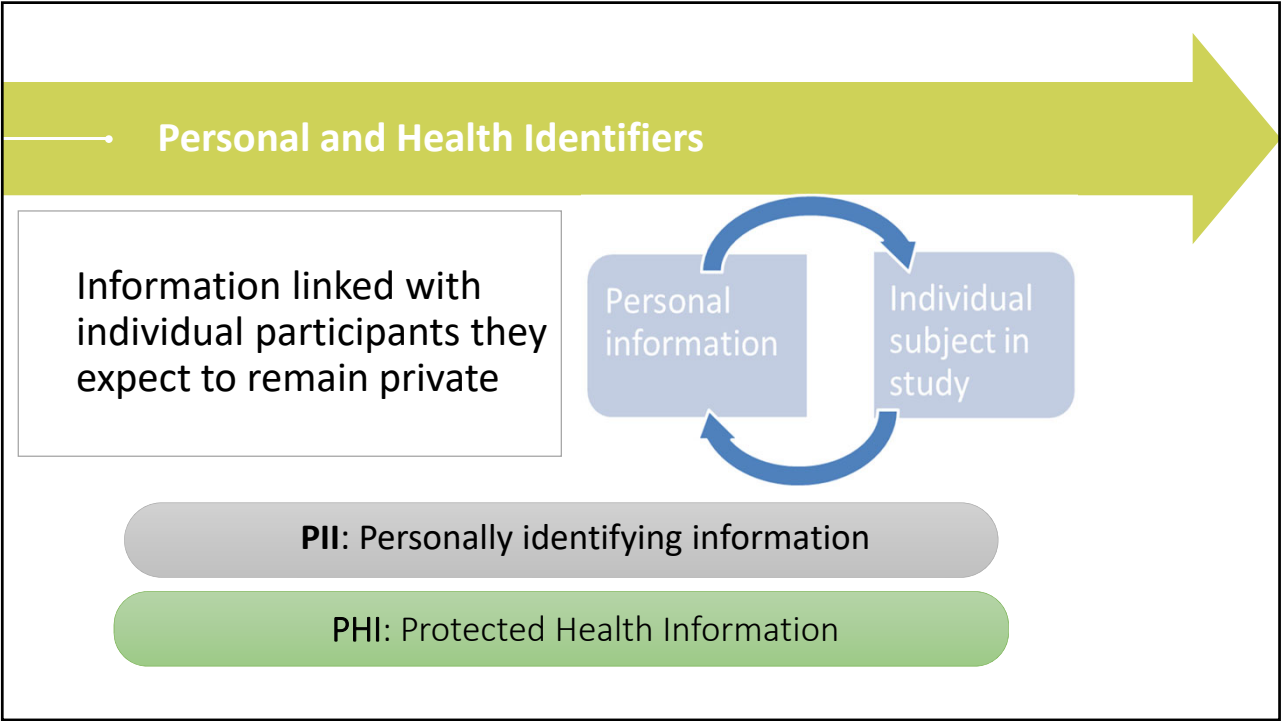
Learning Objectives

- 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)






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

1. Names

3. Dates except year (e.g., birth date, date of research)

4-5. Phone, Fax No.

6. Email addresses

7. Social Security Numbers

8-13. Medical & account numbers, licenses, vehicle/device numbers

14-15 URLs, IP addresses

Less obvious


2. Geographic division smaller than State (e.g. census tract)

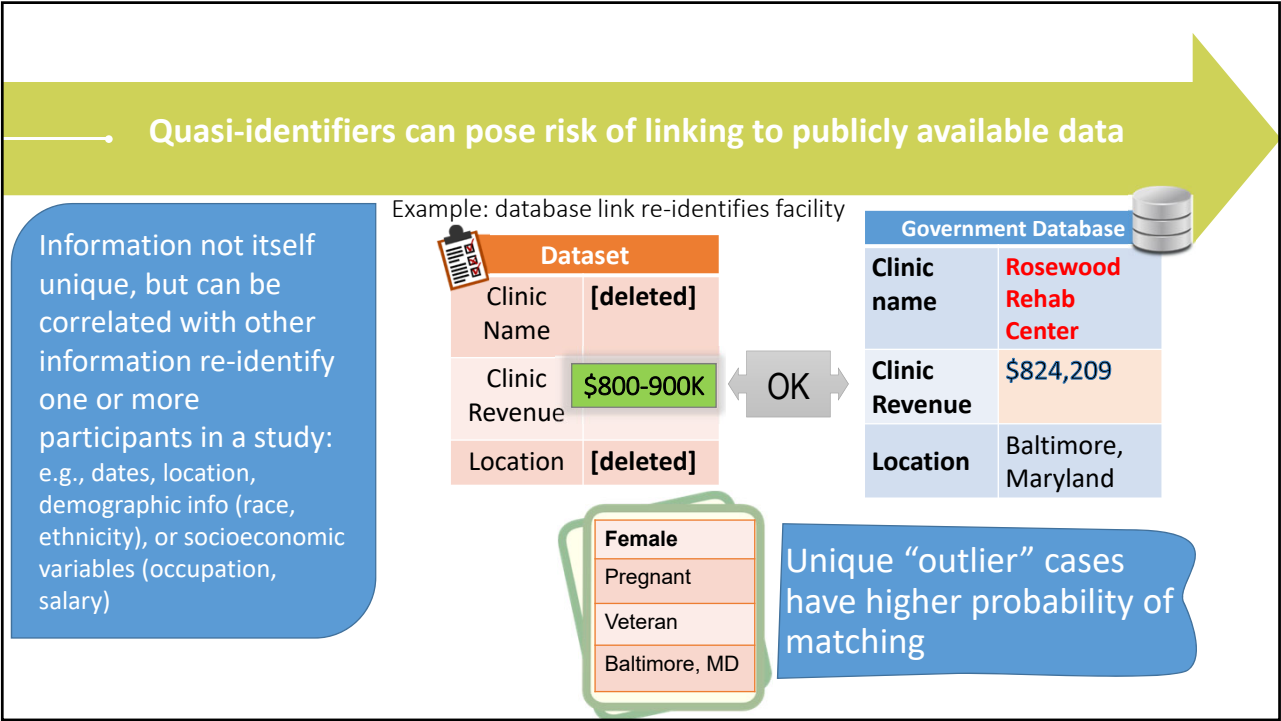
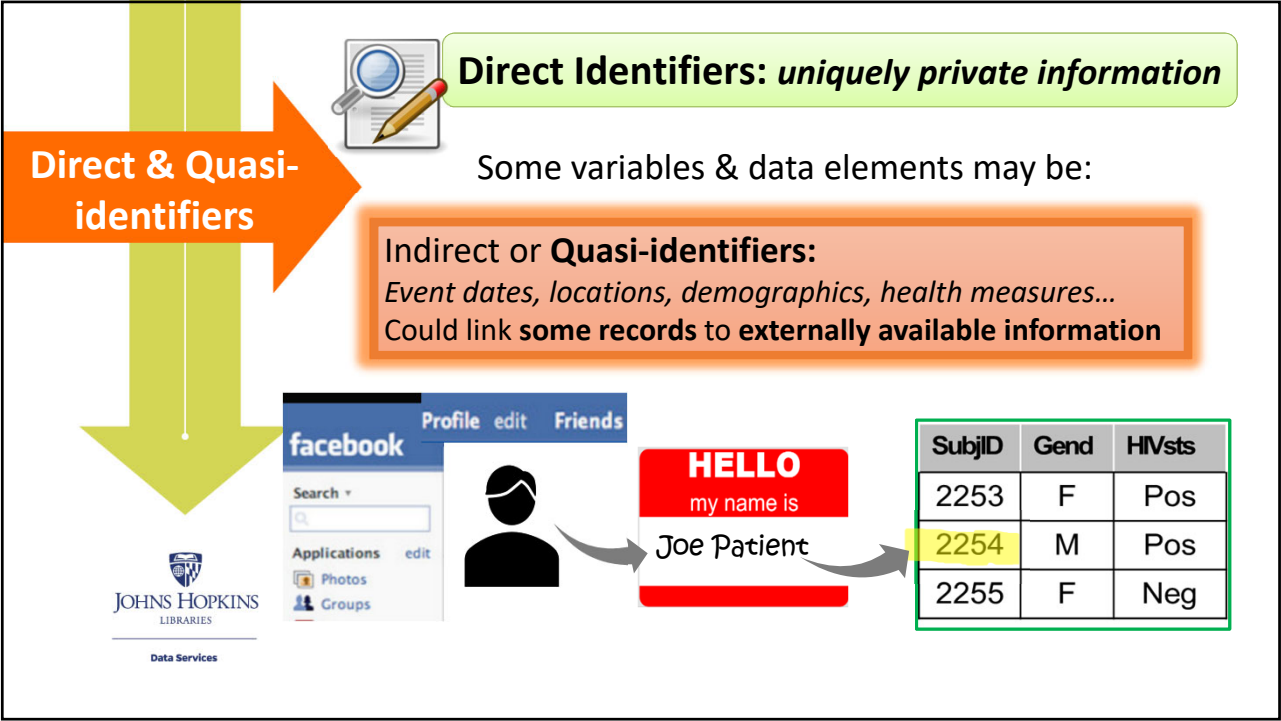
16. Biometric identifiers (fingerprint, voice recordings)

17. Face photos or comparable images

18. Any other unique identifying number, characteristic, or code

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





How to de-identify

De-identifying data to reduce privacy disclosure risk

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

Name	[REDACTED]
Second name	[REDACTED]
Initials	[REDACTED]
Address	[REDACTED]
Tel	[REDACTED]
Email	[REDACTED]

Female
Pregnant
Veteran
Baltimore, MD

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?

How to de-identify

De-identification varies in difficulty

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


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**



BERKMAN CENTER FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY



Tastes, Ties, and Time: Facebook data release
September 25, 2008
In collaboration with Harvard sociology graduate students Kevin Lewis and Marco Gonzalez, and with UCLA

Inferential disclosure	Linkable identifiers in <i>codebook</i> : 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 <i>outliers</i> : 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-identify 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

Proposal justification for restricted access:
"Subject identifiers cannot be adequately removed from all data due in particular to the geographic specificity of the subject population and potential for indirect links to identifiers in external data sources. Selected de-identified datasets will be shared with restricted access at ICPSR"



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

SOM IRB



Research Administration



• IRB requires plans for protecting privacy for shared data.

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

• Data Use Agreements for external collaborations


• May require partial de-identification of shared data


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

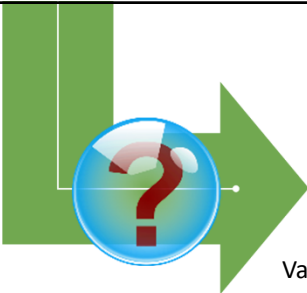
- Protocols are reviewed by the CCDA





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

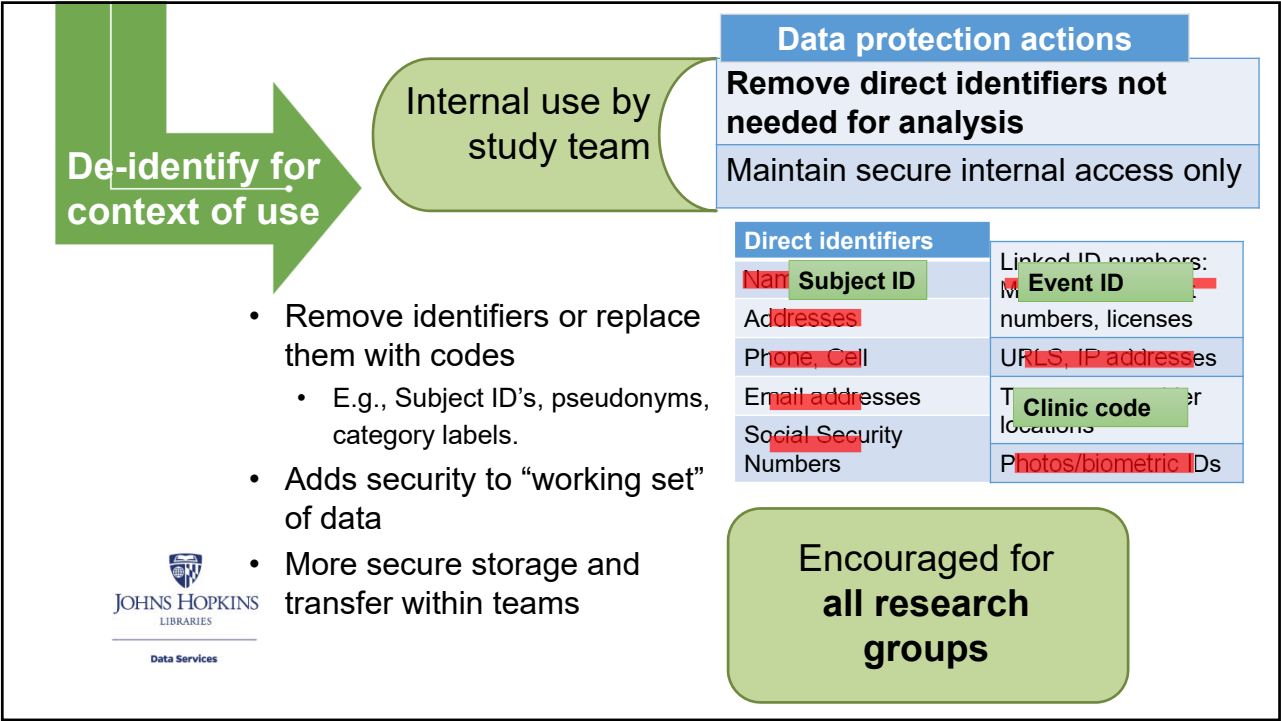
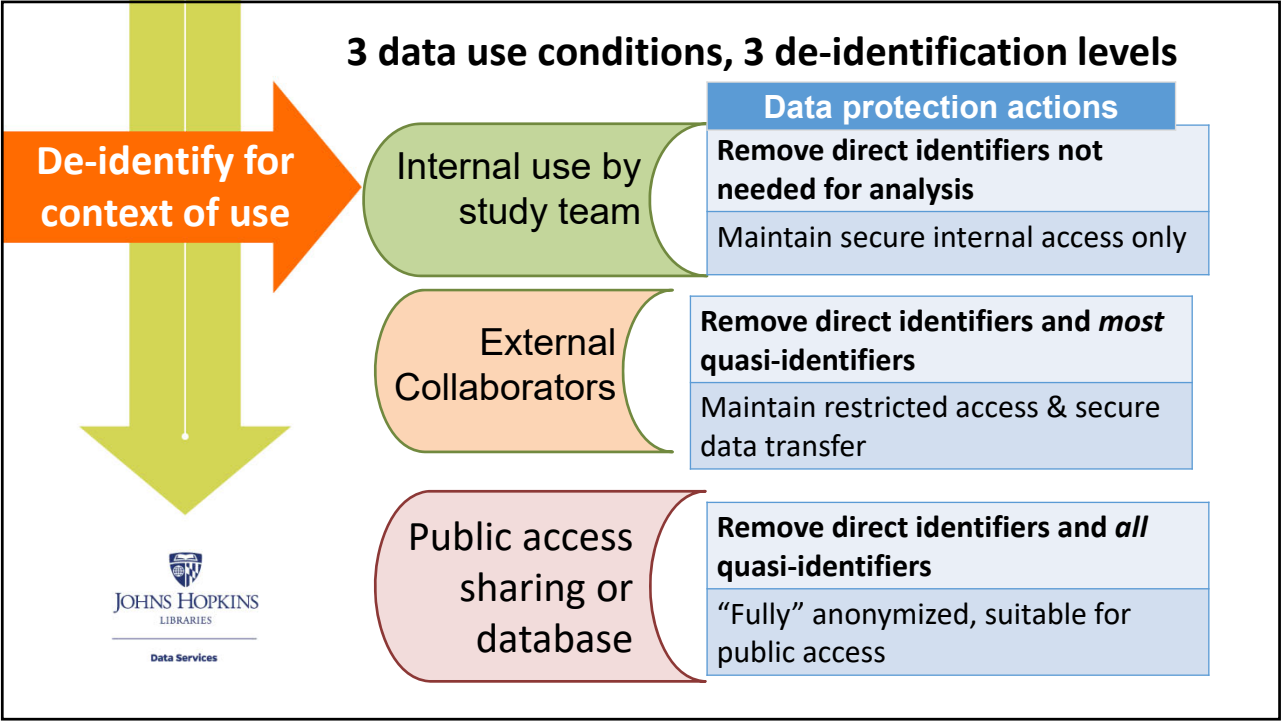
- Sometimes consists only of **masking**, where if key to the original data is known, information can be restored

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







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


Preferred secure platform for PII/PHI

SAFE Desktop: Secure Analytic Framework

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



- 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



De-identify for context of use


External Collaborators

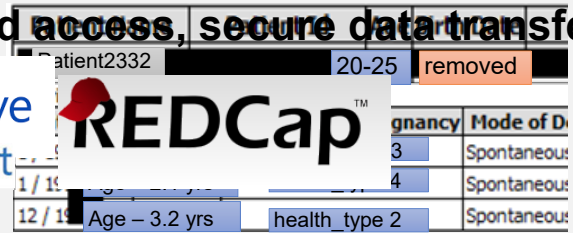
Data protection actions

Remove direct identifiers and most quasi-identifiers


Maintain restricted access & secure data transfer

Requires restricted access, secure data transfer





- Broaden specific values for quasi-identifiers
 - E.g., Numerical values to ranges: Age 52 → 50-55
 - Specific towns to categories (urban/rural)
- Share with Data Use Agreements with IRB-approved researchers



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



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



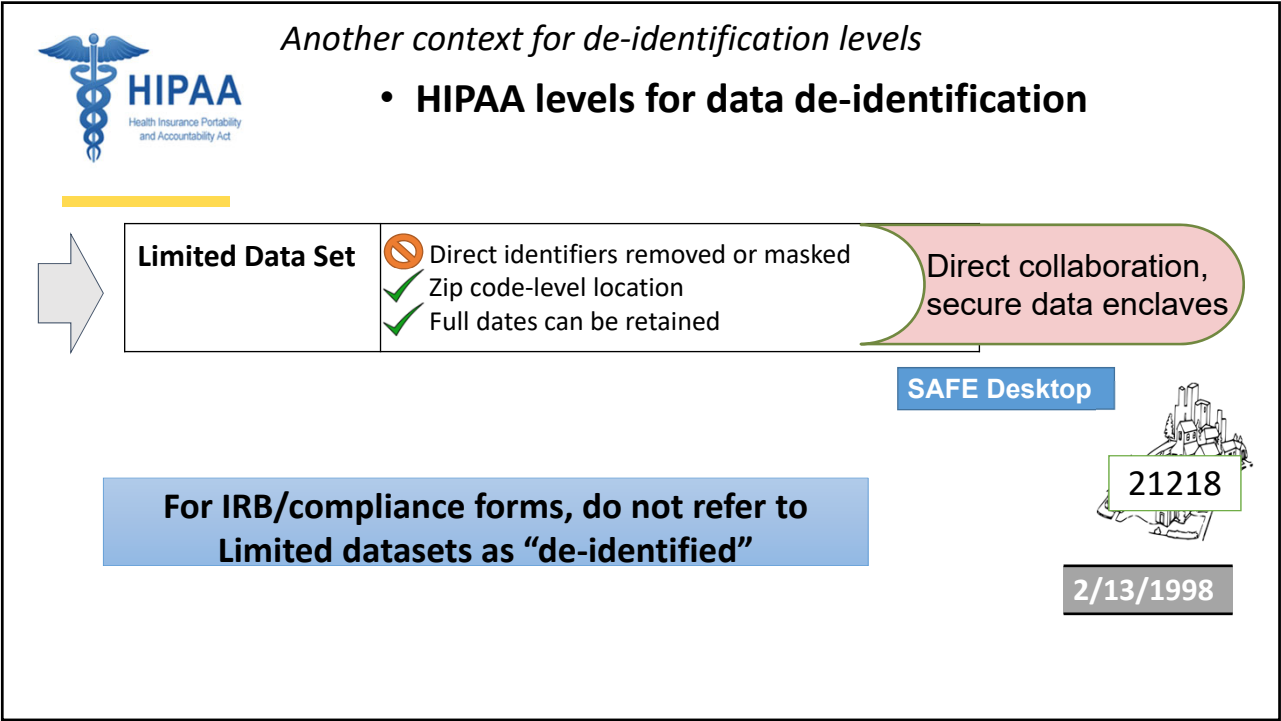
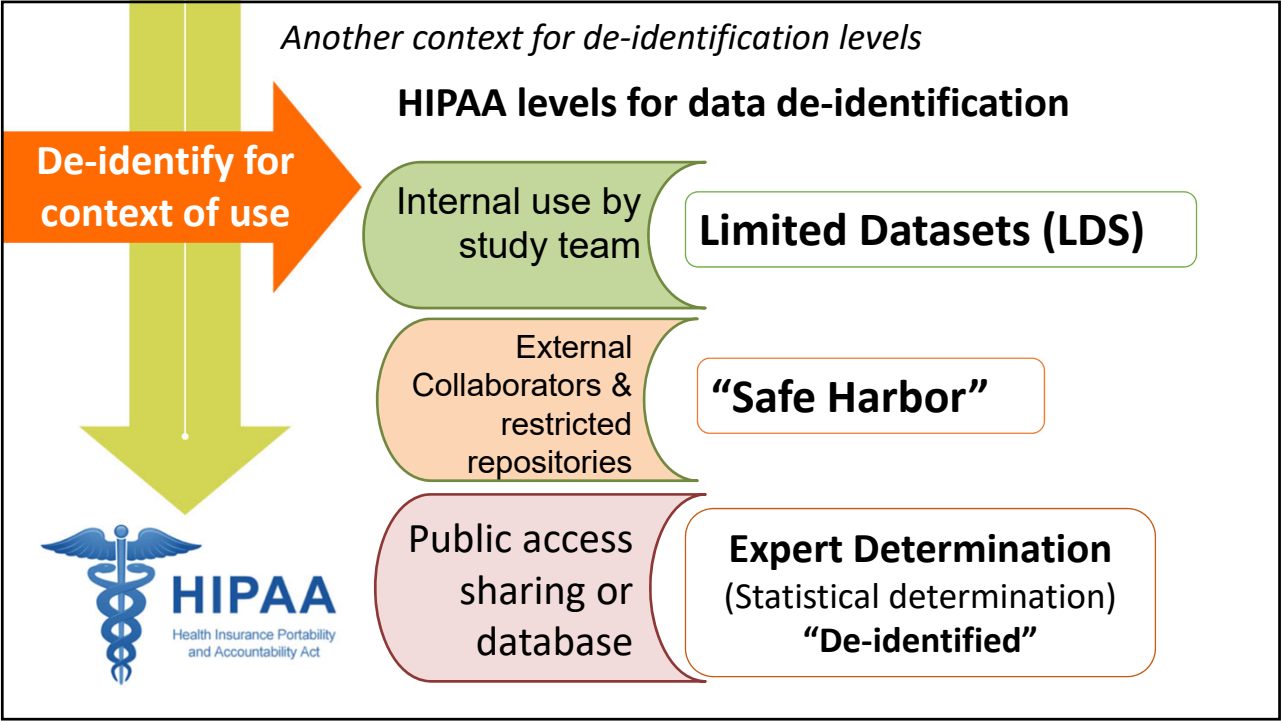
JHU Data Archive


http://archive_data.jhu.edu

Contact dataservices@jhu.edu to archive data.

Do not distribute beyond JHU affiliates without permission. © 2022

15








Another context for de-identification levels

- HIPAA levels for data de-identification**


➔	Limited Data Set	<ul style="list-style-type: none">❌ Direct identifiers removed or masked✅ Zip code-level location✅ Full dates can be retained	Direct collaboration, secure data enclaves
	"Safe Harbor"	<ul style="list-style-type: none">❌ Direct & Quasi-identifiers that "knowingly" re-identify✅ 3-digit Zip code truncation✅ Year only dates, ages above 90	Restricted access repositories


Requires restricted access, secure data transfer



21218

Birth Date	Age
2/13/1927	94 90+ 



Another context for de-identification levels

- HIPAA levels for data de-identification**

➔	Limited Data Set	<ul style="list-style-type: none">❌ Direct identifiers removed or masked✅ Zip code-level location✅ Full dates can be retained	Direct collaboration, secure data enclaves
	"Safe Harbor"	<ul style="list-style-type: none">❌ Direct & Quasi-identifiers that "knowingly" re-identify✅ 3-digit Zip code truncation✅ Year only dates, ages above 90	Restricted access repositories
	Expert/Statistical Determination = "De-identified"	<ul style="list-style-type: none">✅ De-identification performed with appropriate knowledge, accepted statistical techniques✅ Data assessed for remaining disclosure risk✅ Documented methods and results of analysis	Public/open access repository

May require JHM Data Trust approval

Do not distribute beyond JHU affiliates without permission. © 2022

17

• 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

Name


Second name

Initials

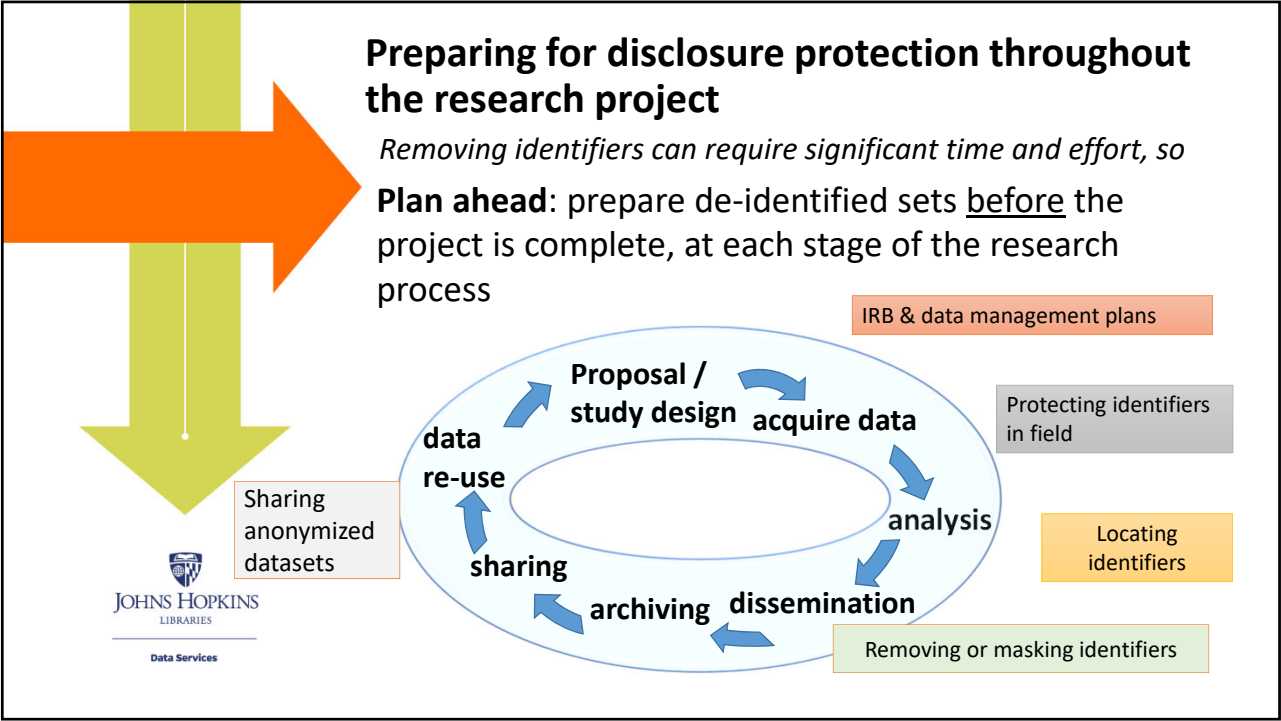
Address

Tel

Email




Preparing for disclosure protection throughout the research project




Plan ahead: disclosure protection and sharing in IRB Forms

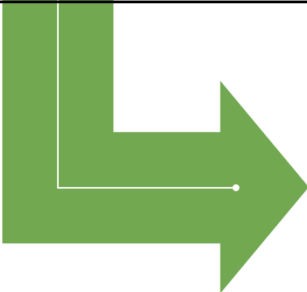
There are two places in your IRB forms where data sharing plans should be addressed:

Research Plan




Consent Form






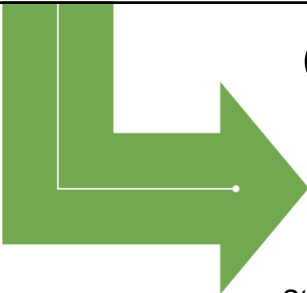
IRB Research Plan (eIRB)

- **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?*



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








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:

Too vague:	Your de-identified data will be shared in a publicly accessible data archive.







Consent Form Language

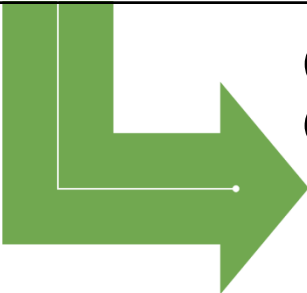
- In most cases, it is essential to get participant's consent to share data online in a repository, even for restricted repositories and for most de-identified datasets

Statement examples:

State <u>where</u> de-identified data will be shared:	Your responses will not be linked to your private information, but selected anonymized 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 <u>small</u> 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.





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

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

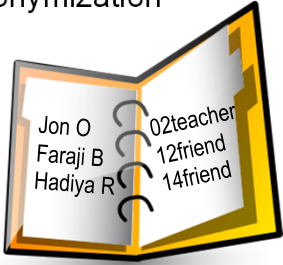

Protect Identifiers During Data Collection

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?



Do not distribute beyond JHU affiliates without permission. © 2022

22


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
<https://ictr.johnshopkins.edu/taq/safe-desktop/>


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


 **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.

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


 Never Dropbox!

Analysis phase: marking & changing identifiers


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

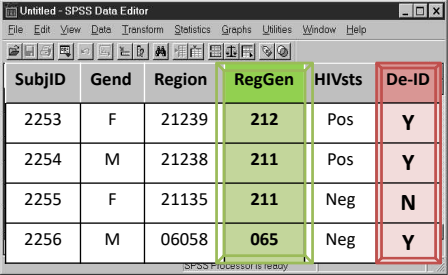
When you see an identifier, change it or tag it to change later


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



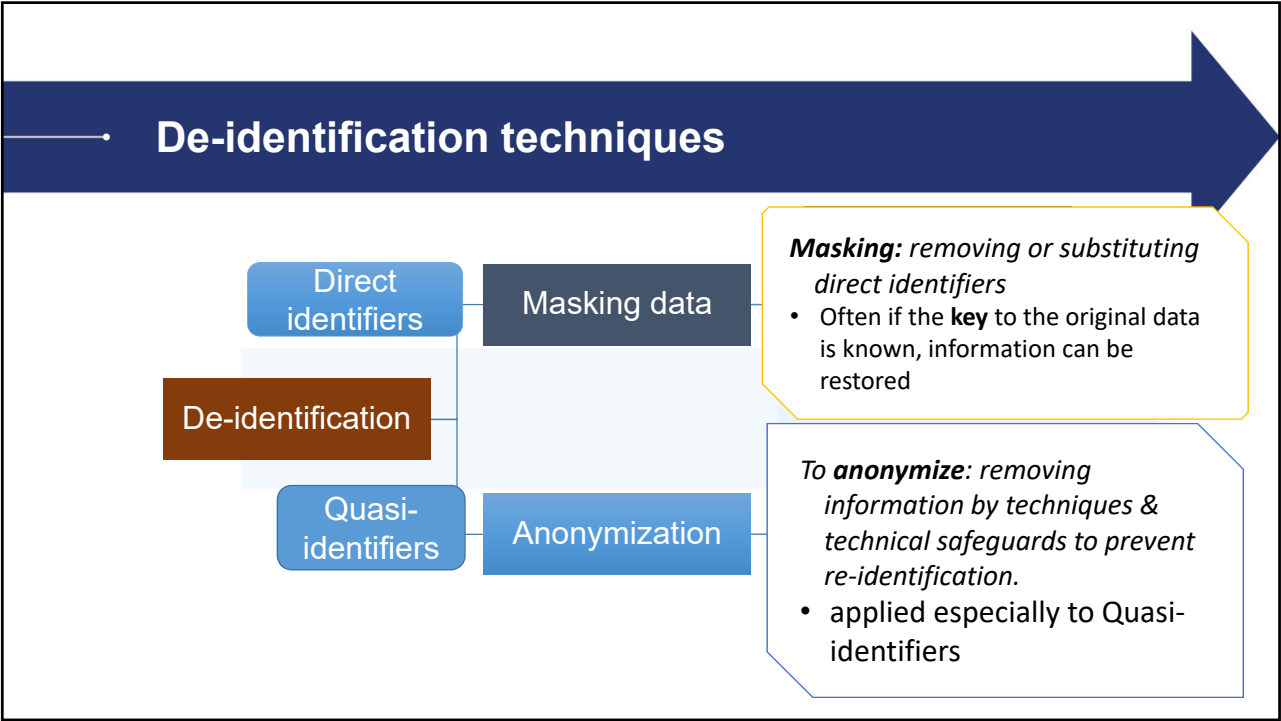
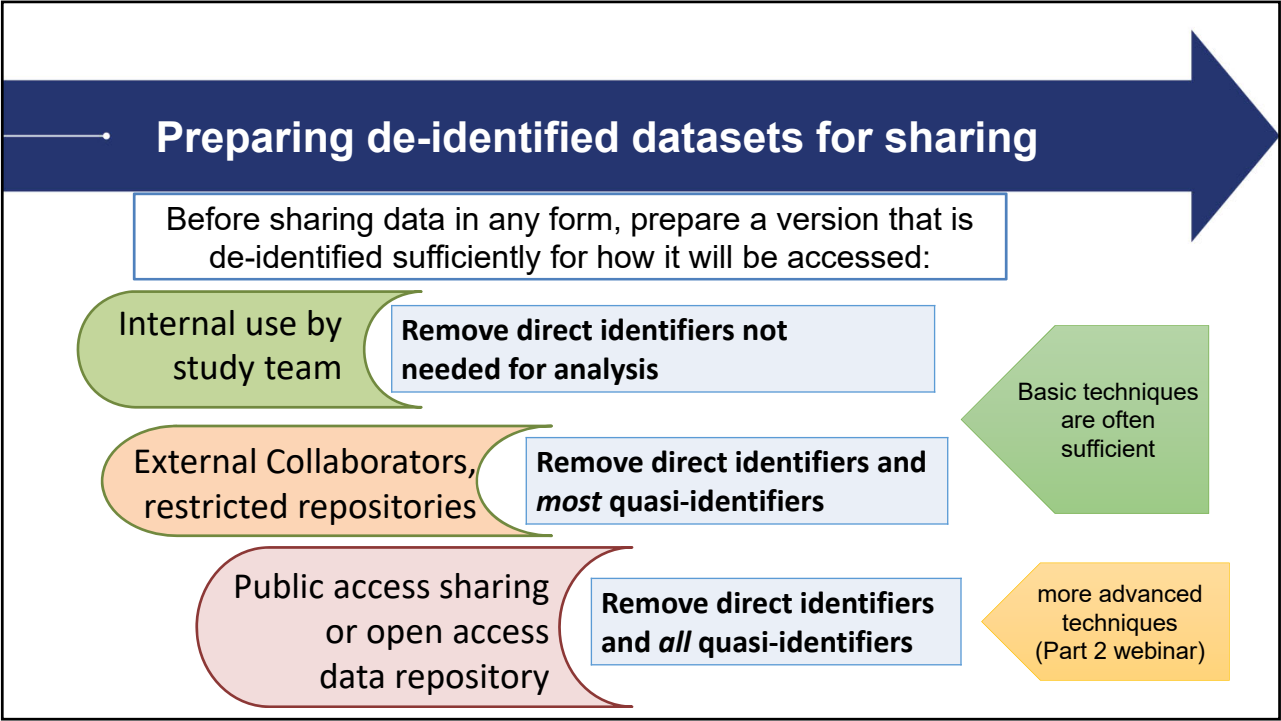


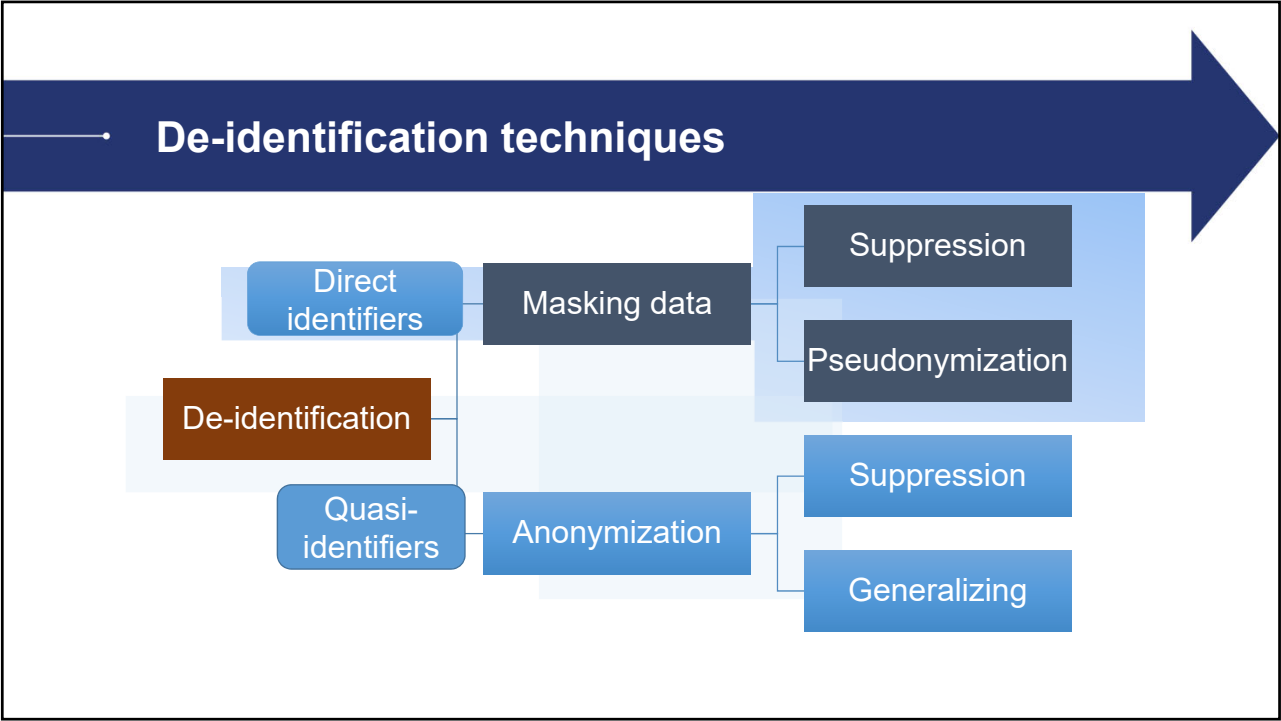


Preparing de-identified datasets for sharing

Do not distribute beyond JHU affiliates without permission. © 2022

24





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: B2-2296		Report Date: 2006-								
Patient Name	Patient Id	Age/Birth Date	Address	Husband	Clinic Name					
Patient2332			married	Inst_OC3						
Lab Tests		Date Request		Lab Type	Date Results	Results	Site	Clinician	Drug Interventions	
		2006-	Hb - 1st screen	2006-			InstOC3	Dr_16	2006-	X X

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

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

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			W	0	83
495.4649		45	W	0	
384.5498		33	W	0	30
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**

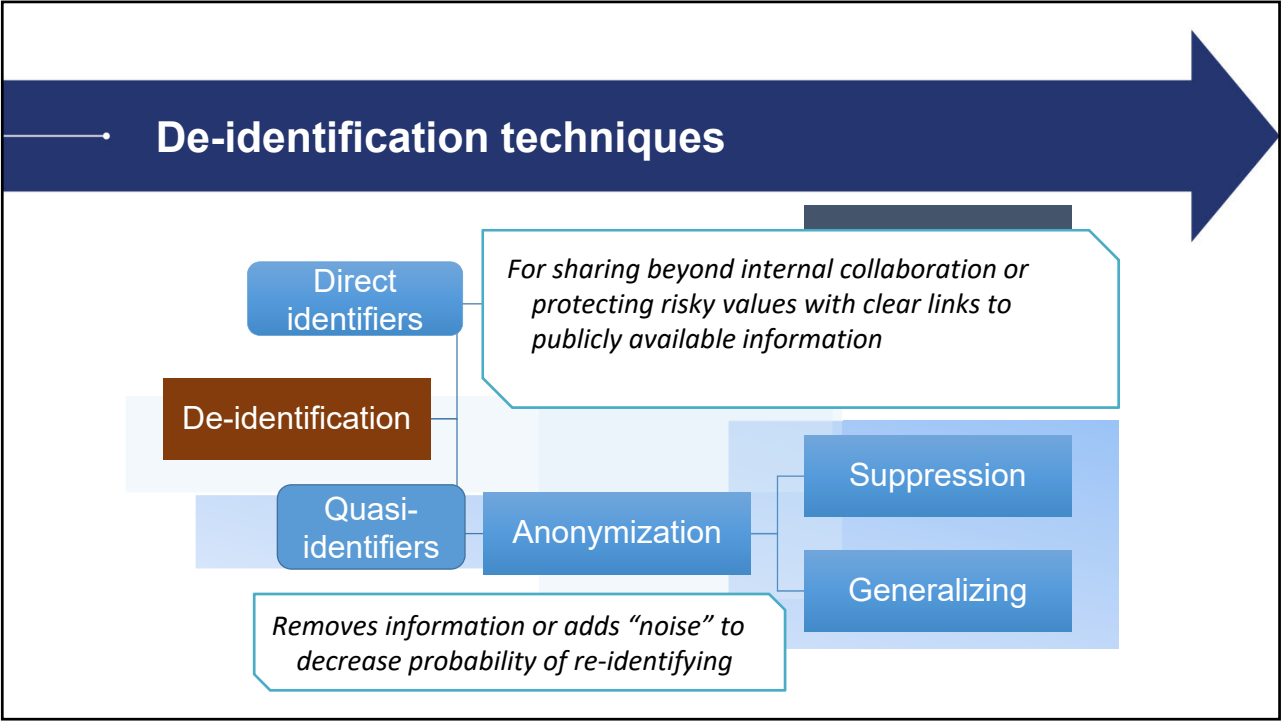
MRN	Last Name	CSN
17274.685	Mathews	924
16303.342	Jones	456
18048.841	McGaren	333

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

Test	[table3]
CSN	456
BP	120/60

- 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.

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



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

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

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

Do not distribute beyond JHU affiliates without permission. © 2022

28

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

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

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

Restricted access among collaborators:

Remove direct addresses, generalize Zip code to 3 digits

For full de-identification:

Areas ≥ State/province

Categorize regions: urban/rural

population blocks ≥ 100,000

Advanced: GIS polygon mapping

Postal code:

21219

Generalization of other data type

Top and bottom coding:

• change extreme top & bottom of outlier variables:

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

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+

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

Change: Recode revenue as range or grouped averages

Remove potential links of no analytic value

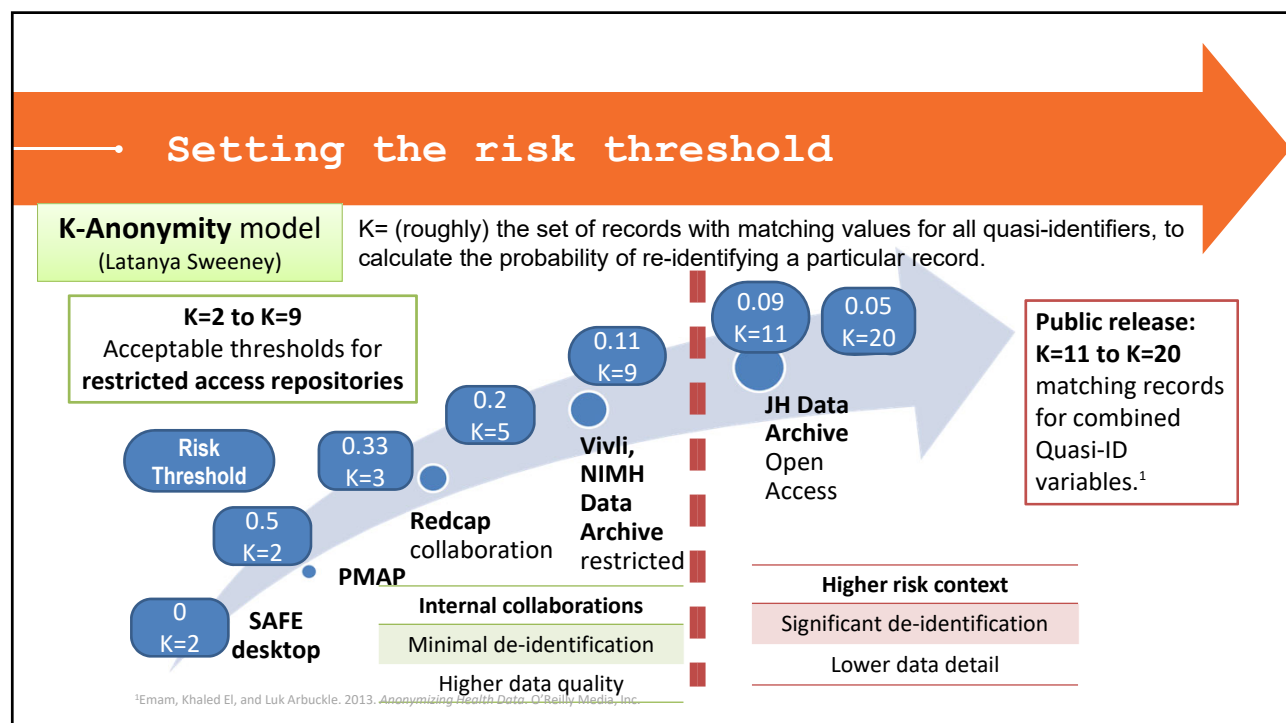
Weigh risks/utility of removing/recoding based on likelihood of external link

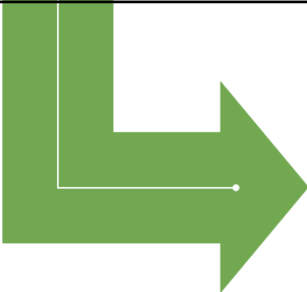
JOHNS HOPKINS LIBRARIES

Data Services

Do not distribute beyond JHU affiliates without permission. © 2022

30






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 **Nov 8 @ 12:30 pm – 2:00 pm** **Online**

$$h(X) = -\int_0^1 0.5 \log_2 0.5 \, dx - \int_{-\infty}^{\infty} 0.5 \log_2 0.5 e^{-0.5 a^2} \, da$$
$$\int_0^1 0.5 \log_2 0.5 \, dx = -0.5 \log_2 0.5$$
$$I(A|B) = 1 - \Pi(A|B)/\Pi(A) = 1 - 2^{h(A|B)}/2^{h(A)} = 1 - 2^{-I(A;B)}$$



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



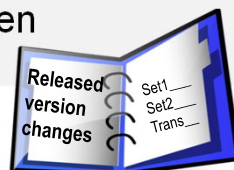
Documenting & closing projects

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.


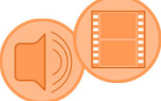




Can software de-identify for us?


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-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...]



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)



