



All about Data

Meeting at Faculty of Tropical Medicine, Mahidol University

24 May 2017



Data

- ❖ Data collection
- ❖ Data banking
- ❖ Data sharing
- ❖ Data mining

Ethical Guidelines
Council for International
Organization of Medical Science
(CIOMS)



CIOMS guidelines -Collection, storage and use of

Guideline 11: Biologic material and related data

Guideline 12 Data in health-related research

- the vast majority of people **do not object** to their biologic material & data being stored in collections and used for **research for the common good**
- the precise nature of the **future research** is typically **unknown**, it is **impossible to obtain specific** informed consent at the time of collection.



Broad consent or Informed opt-out



Commentary on Guideline 11

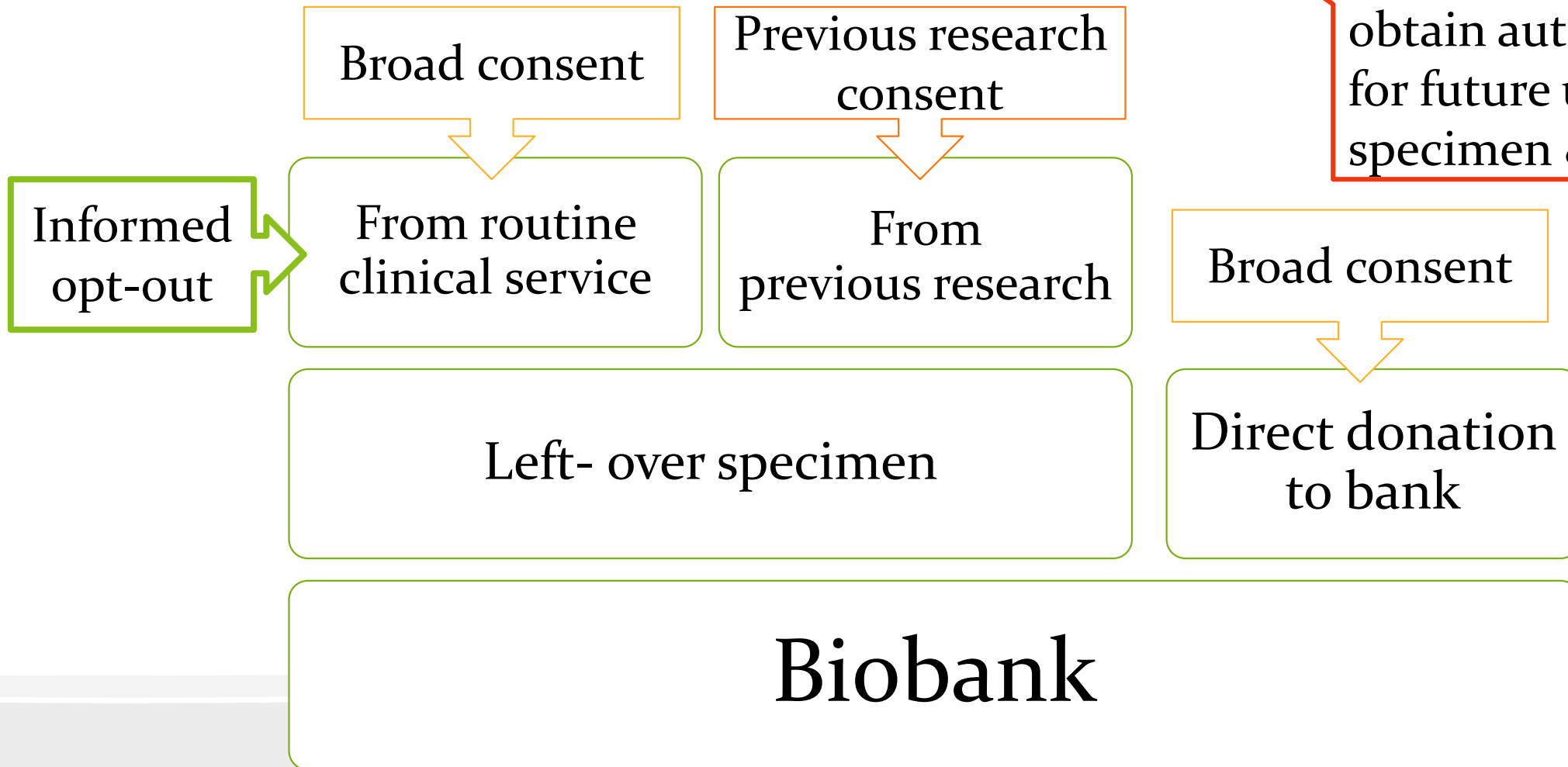
- **Human biological materials** may include:
 - tissues, organs,
 - blood, plasma, serum,
 - DNA, RNA, proteins,
 - cells, hair, nail clippings, skin,
 - urine, saliva, or other bodily fluids
- **Source**
 - diagnostic or therapeutic procedures,
 - autopsy specimens,
 - donations of organs or tissue from living or dead humans,
 - bodily wastes or abandoned tissue

CIOMS guidelines -Collection, storage and use of

Guideline 11: biologic material and related data

Guideline 12 Data in health-related research

institutions must have a governance system to obtain authorization for future use of these specimen & data





CIOMS Guideline 11 Biospecimen & related data

VS 12 Data in health-related research

institutions must have a governance system to obtain authorization for future use of these data

- When *specimens* are collected for research purposes, either specific informed consent for a particular use or broad informed consent for unspecified future use must be obtained from the person from whom the material originally is obtained.
- When human *biological materials* are **left over after clinical diagnosis or treatment** (so-called “residual tissue”) and are stored for future research, a specific or **broad informed consent** may be used or may be **substituted by an informed opt-out procedure**.
- When *data* are collected and stored for research purposes, either specific informed consent for a particular use or broad informed consent for unspecified future use must be obtained from the person from whom the data were originally obtained.
- When *data* are used that were collected in the context of **routine clinical care**, an **informed opt-out procedure must be used**.
- This means that the *data* may be stored and used for research unless a person explicitly objects.
- However, a person’s objection is not applicable when it is **mandatory** to include *data* in **population-based registries**.



Informed Opt-out

the material is stored and used for research unless the person from whom it originates explicitly objects.

The informed opt-out procedure must fulfil the following conditions:

- 1) patients need to be aware of its existence;
- 2) sufficient information needs to be provided;
- 3) patients need to be told that they can withdraw their data;
- 4) a genuine possibility to object has to be offered.



Broad consent: Biospecimen VS. Data

- Broad informed consent is **not blanket consent** that would allow future use of bodily material **without any restriction**.
- On the contrary, broad informed consent **places certain limitations** on the future use of bodily materials.
- **Secondary use of stored data:** collected in databanks, during research or during other activities (for example, clinical practice, health insurance)
- In those cases, it is acceptable to use the data for secondary analysis when the intended use **falls within the scope of the original** (broad) informed consent



Broad Consent

- the **purpose** of the databank;
- the **conditions and duration** of storage;
- the **rules of access** to the databank,
- the ways in which the **donor can contact** the databank custodian and remain informed about future use;
- the **foreseeable uses** of the data, whether limited to an already fully defined study or extending to a number of wholly or partially undefined studies;
- **who will manage** access to the data;
- the intended goal of such use, whether **only for basic or applied research**, or also **for commercial purposes**;
- the possibility of **unsolicited findings and how they will be dealt with**.

Broad consent: Biospecimen VS. Data

Biospecimen

- the purpose of the [biobank](#);
- the conditions and duration of storage;
- the rules of access to the [biobank](#);
- the ways in which the donor can contact the [biobank](#) custodian and remain informed about future use;
- the foreseeable uses of the [materials](#), whether limited to an already fully defined study or extending to a number of wholly or partially undefined studies;
- the intended goal of such use, whether only for basic or applied research , or also for commercial purposes; and
- the possibility of unsolicited findings and how they will be dealt with

Data

- the purpose of the [databank](#);
- the conditions and duration of storage;
- the rules of access to the [databank](#),
- the ways in which the donor can contact the [databank](#) custodian and remain informed about future use;
- the foreseeable uses of the [data](#), whether limited to an already fully defined study or extending to a number of wholly or partially undefined studies;
- **who will manage access to the data**;
- the intended goal of such use, whether only for basic or applied research, or also for commercial purposes;
- the possibility of unsolicited findings and how they will be dealt with.



Research ethics committees and biobanks/Databank

- The protocol for **every study must be submitted** to a research ethics committee
 - To ensure that the **proposed use** of the materials **falls within** the scope specifically agreed to by the donor - **broad informed consent for future research**
 - Determination for **re-consent** If the proposed use falls outside the authorized scope of research
 - May **waive** the requirement of individual informed consent for research if satisfy the predetermined conditions



Waiver

- When researchers seek to use stored materials collected for past research, clinical or other purposes without having obtained informed consent for their future use for research, the research ethics committee may waive the requirement of individual informed consent if:
 - 1) the research **would not be feasible or practicable to carry out without the waiver**;
 - 2) the research has ***important social value***; and
 - 3) the research poses ***no more than minimal risks*** to participants or to the group to which the participant belongs.



Special concerns for using Data/Data banking

- Re-contacting participants
- Data mining
- Limits of confidentiality
- Data sharing - [refer to guideline 24](#)



Data sharing- guideline 24

- Requires careful balancing of competing considerations
- Researchers must **respect the privacy and consent** of study participants
- Funders and sponsors must **require** funded researchers **to share** study data and must **provide appropriate support** for sharing
- Research institutions and universities must **encourage** researchers **to share** data.
- The risks of data sharing may be mitigated by controlling **with whom** the data are shared and **under what conditions**, without compromising the scientific usefulness of the shared data.
- **Organizations** that share data should **employ data use agreements**, observe additional **privacy protections beyond de-identification** and data security, as appropriate, and appoint an independent panel that includes members of the public to **review data requests**.
- These safeguards **must not unduly impede access** to data.

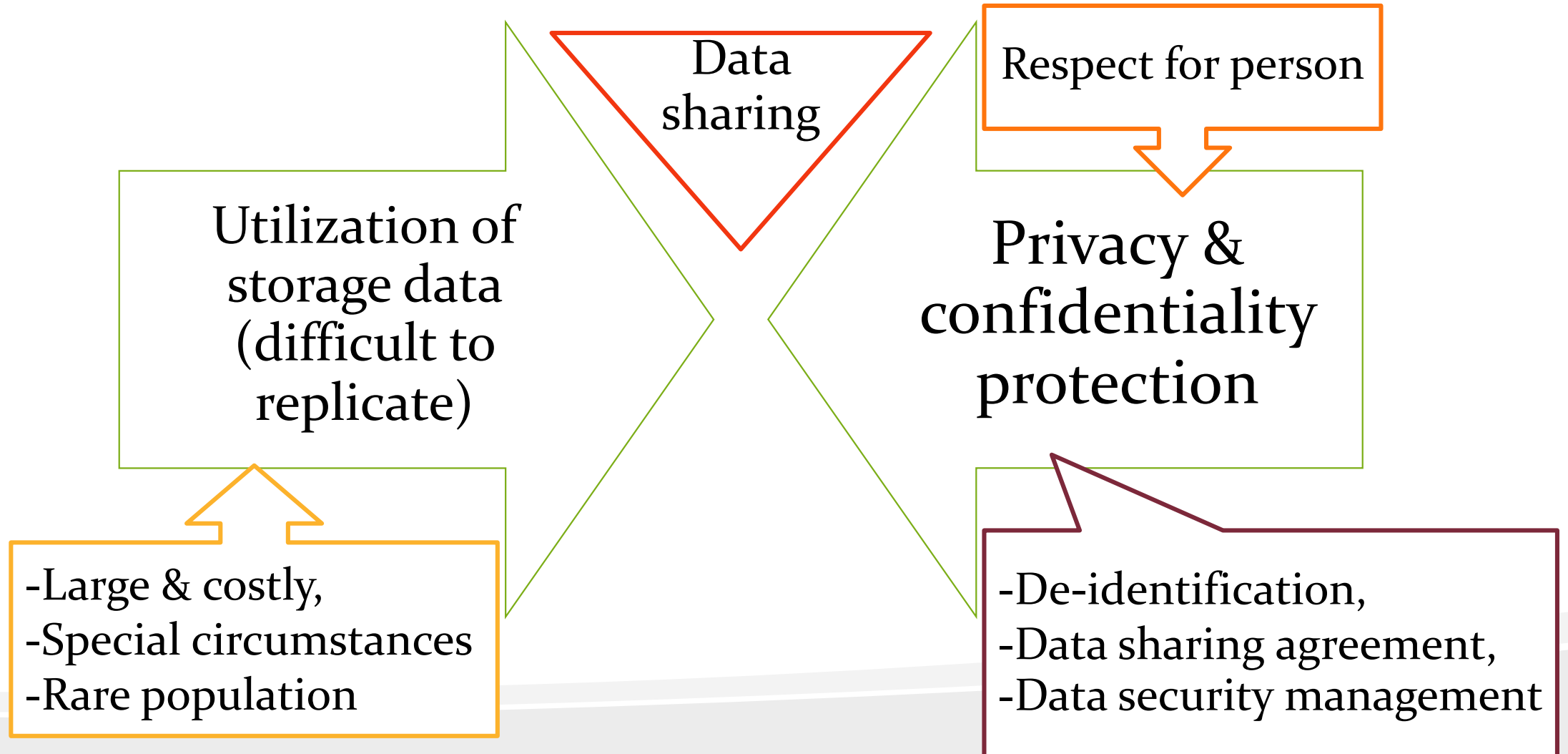


Overview

- Publishing the raw data constitutes an acceptable mechanism for sharing data,
- Raw data from large studies are not amenable to sharing through publication.
- The **rights and privacy** of individuals who participate in research **must be protected at all times**, patentable and other proprietary data should also be protected.



Overview

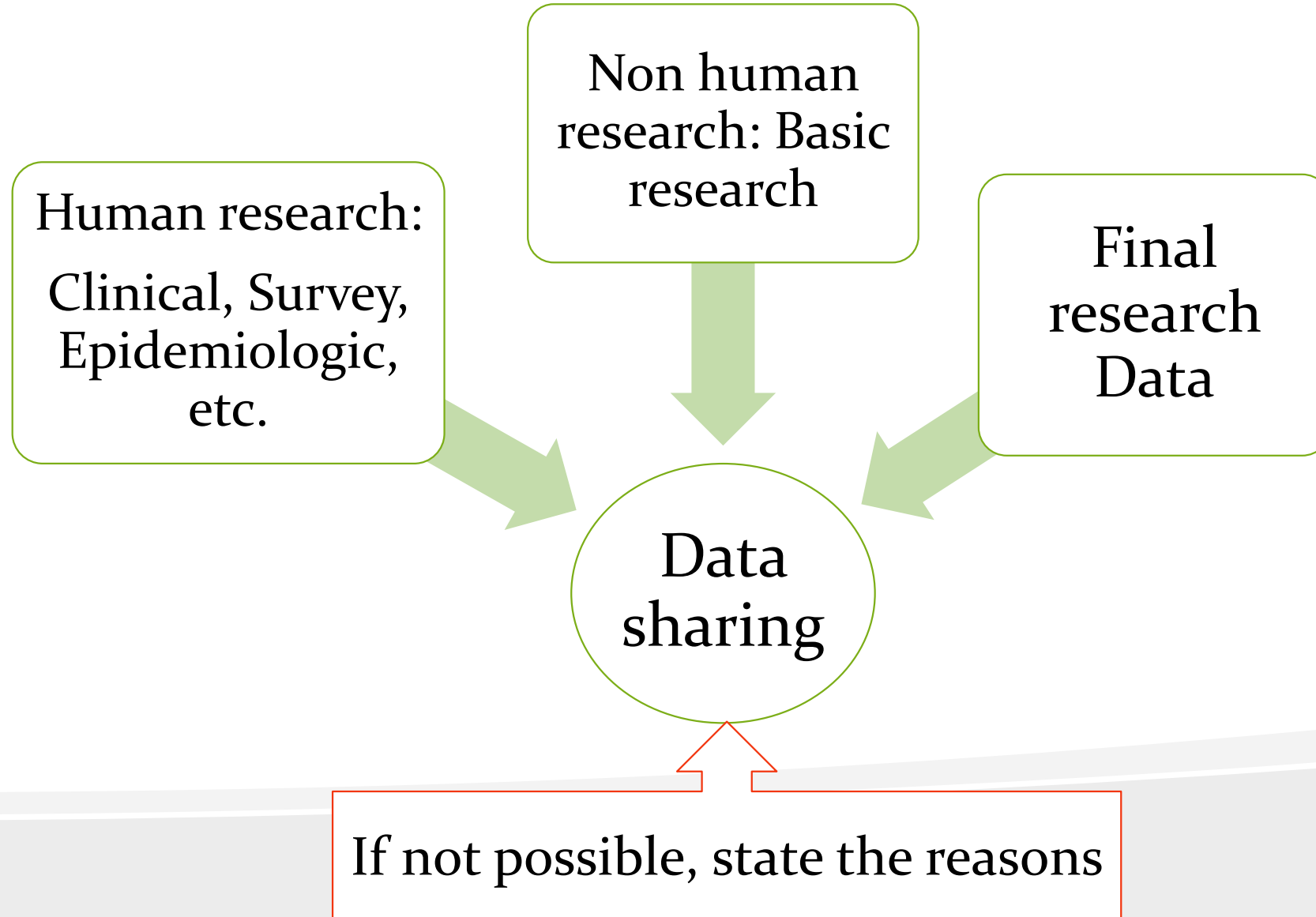




Overview

- Data that constitute "**unique resources**" especially should be shared unless there is a strong reason not to.
- Such data are **difficult if not impossible to replicate** because of
 - cost (e.g., large national longitudinal surveys),
 - special circumstances (e.g., health effects associated with a natural disaster),
 - rare population (e.g., a sample of centenarians)
- **Less likely candidates** for sharing are data from **small studies** involving research procedures that are **easily replicated** or data from **human subjects that might identify** them.

Applicability





Data Sharing Across Countries

U.S. investigators
collecting data in
other countries

Investigators
from *foreign*
institutions
(work with US)

- familiarize themselves with the policies governing data sharing in the **countries in which they plan to work**
- address any specific **limitations** in the data-sharing plan in their application



Data Documentation

- Proper documentation is needed to ensure that **others can use the dataset** and to **prevent misuse, misinterpretation, and confusion**.
- Documentation should provide information
 - the **methodology and procedures used to collect** the data,
 - details about **codes**,
 - definitions of **variables**, variable field locations, frequencies, and the like.
- The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset.




Data Documentation

- It is appropriate for scientific **authors** to **acknowledge the source of data** upon which their manuscript is based.
- It could be in methods and/or reference, or in acknowledgement sections of the manuscripts,
- Authors using shared data should check the policies of the journal to which they plan to submit to determine the precise location in the manuscript for such acknowledgement.
- Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication.



Timeline

- The Office for Human Research Protections (OHRP) requires **research records** to be **retained** for at least 3 years after the completion of the research.
- Any research that involved collecting **identifiable health information records** must be **retained** for a minimum of **6 years** after each subject signed an authorization.



Data sharing should occur no later than the acceptance for publication of the main findings from the final dataset.



Human Subjects and Privacy Issues

- The **PI, IRB**, and the **Institution** have **responsibility** to protect the rights of subjects and the confidentiality of the data.
- Prior to sharing the data,
 - **de-identify** the data.
 - **removing indirect identifiers** and other information that could lead to "deductive disclosure" of participants' identities.



study design,
the informed consent documents,
the structure of the resulting dataset



Human Subjects and Privacy Issues

- Researchers who seek **access** to individual level data are typically required to enter into a **data-sharing agreement**.
- Researchers who are planning clinical trials and **intend to share** the resulting data should think carefully prior to the initiation of the study
 - the study design,
 - the **informed consent** documents,
 - the structure of the resulting dataset



HIPAA Privacy Rule, de-identification of a dataset means removing the following variables:

- Names; Social Security Number; medical record and prescription numbers; health plan beneficiary number; account numbers; certificate or license number;
- Geographic information (including city, state, and zip code);
- Elements of dates such as those for birth, hospital admission and discharge, death;
- Telephone numbers; fax numbers; Web Universal Resource Locator (URL); Internet Protocol (IP) address number; electronic mail addresses;
- Any vehicle identifier or serial number, including license plate number; any device identifier or serial number;
- Any biometric identifiers, including finger or voice prints; full face photographic images or any comparable images; and
- Any other unique identifying number, characteristic, or code consisting of any segments of the previously listed identifiers.



Personal identifiable data

Direct
information:
name, ID numbers

Biometric
identifiers: finger
print, photograph,
images, etc.

Telephone
number, IP
address, URL,
email address

Related date: birth,
hospital
admission, etc.

Vehicle & device
identifiers

Geographic
information: home
& office address,
zip code



Measures used to minimize the risk of breaching the confidentiality of data include the following:

- Mandatory agreements to maintain confidentiality
- Data encryption
- Electronic firewalls and locked storage facilities,
- Password authentication of users
- Audit trails
- Disaster prevention and recovery plans
- Security measures for backup tapes.

Institutions and investigators should work closely to develop and update plans and procedures to protect the security of data.



Data Sharing- Policy & Procedures

- Data Archive
- Data Enclaves
- Mixed mode sharing



Data Archive & Enclave

- **Data Archive** - A **place** where **machine-readable data** are acquired, manipulated, documented, and finally distributed to the scientific community for further analysis.
- **Data Enclave** - A **controlled, secure environment** in which eligible researchers can **perform analyses using restricted data resources**.
 - Datasets **cannot be distributed** to the general public
 - Use agreements that **prohibit redistribution**



Mixed Mode Data Sharing

- This method allows for **more than one version** of the dataset and provides **different levels of access** depending on the version.
- For example, a **de-identified** dataset could be made available for **general use**, but **stricter controls** through a data enclave would be applied if access to **more sensitive data** were required.



Data-sharing agreement

- to impose appropriate **limitations** on users.
- an agreement usually indicates
 - the criteria for data access, whether or not there are any **conditions** for research use,
 - incorporate privacy and confidentiality standards
 - to ensure **data security at the recipient site** and
 - **prohibit manipulation** of data for the purposes of identifying subjects.



Data-use sharing agreements (also known by other names:- license agreements, data-distribution agreements, and data-sharing agreements)

- who can use the data
- how they are to be used
- the **privacy** and the **confidentiality protection**
- incorporate confidentiality standards to ensure **data security at the recipient site**
- **prohibit manipulation** of data for the purposes **of identifying** subjects.
- stipulate that the recipient **not transfer** the data to other users,
- the data are **only** to be used **for research** purposes,
- the proposed research using the data will be **reviewed by an IRB**, and the like.
- **Penalties** for violating terms of the agreement are generally specified in these agreements.



Data Archive

Data-Sharing Agreement

- **Individual** user or **Institution**
- If the purchaser is an institution, an **institutional representative must sign an agreement** certifying that
 - Only faculty, students, and staff can use the data.
 - Neither printed nor electronic data may be copied or otherwise shared.
 - Use of the data is restricted to statistical reporting, analysis, and teaching.
 - Prohibits the user from making any efforts to identify individual cases
 - Prohibits linking data from this archive with individually identifiable data from other datasets.
- **Violation** of the license agreement carries **civil liability**.



Restricted access

- The restricted access dataset is available only to certified researchers who provide a nonrefundable fee to cover administrative handling charges and user support.
- Provider **embedded a hidden signature** identifying the purchaser in each electronic file, so that **unauthorized copies can be traced**.



Data Enclaves

- Submit request to the committee
- Run analysis under supervision

- In order to gain access to restricted data, **researchers must submit to the committee**
 - a detailed description of their projects
 - summary of the proposed research including a statement of why publicly available data are insufficient,
 - a complete list of data requested, including data system, files, years, variables, and the like
 - personal identification and institutional affiliation, a current resume
 - source of funding,
- This review addresses the following critical questions:
 - Does the proposed activity constitute statistical research or an illegal attempt to identify respondents?
 - If it is research, is there any risk that respondents will be identified inadvertently?



Data Enclaves

Sign an agreement of confidentiality prohibits
Send manuscript for disclosure limitation review

- All applicants are also required to sign an agreement of confidentiality **prohibits**
 - copying files or portions of files,
 - keeping restricted materials,
 - attempting to learn the identity of participants,
 - removing any print outs, electronic files, or other documents from the enclave unless authorized by provider.
- Upon completion of the project, the recipient **must return**
 - all biomaterials, clinical and genetic data received or
 - certify that the clinical and genetic data were destroyed in accordance with applicable laws and safety procedures
- All papers or reports submitted for publication **must first be submitted to data provider for disclosure limitation review.**



Data Enclaves

- The fee charged for work at the data enclave (\$200 per day or \$1,000 per week) includes
 - space,
 - equipment,
 - staff time for supervision and disclosure limitation review,
 - the creation and maintenance of data files required by the researcher.
- All work must be **completed within the confines** of the enclave.
- Researchers must **work under the supervision** of provider staff during normal working hour
- No electronic or hard copies of data can leave the facility unless they are submitted to a disclosure limitation review.



Protection of subjects

- Separation of identities from the data immediately after data collection.
- Only a Security Manager can link the name and address of respondent to interview data.
- The investigators also asked for and received a Certificate of Confidentiality from DHHS to protect subjects' identities.
- All Health staff are required to take training in data confidentiality and security issues.
- Individuals and institutions seeking to obtain the Health data are encouraged to develop and implement a similar training program.
- Only certified researchers are permitted access to Health data



Protection of subjects

- All users must **sign an agreement** to
 - **maintain privacy** of subjects and **confidentiality** of the data.
 - **complied** with a set of security requirements covering how the data are handled and stored. These requirements are updated periodically to reflect changes in computer technology.
 - **submit** letters from their **IRBs** verifying and **approving plans for data security** and for minimizing risks of deductive disclosure.
- The staff from Health **conducts site visits to monitor the use of these data** at outside institutions. The user fee covers the cost of these visits.
- Researchers requesting use of data that cannot be shared through contractual agreements must come to the Health site to conduct analyses under the supervision of Health staff.



Plan for Data Sharing

Host

- Plan for
 - Data Collection: anonymous, anonymized (de-identify)
 - Data Sharing 3 modes: archive, restricted access, enclave
- Data security management
- Data sharing agreement
- Monitoring
- Staff Training

Recipient: Individual researcher & institute representative

- Training in data confidentiality & security
- IRB approval
- Submit request for Access certification
- Sign Data sharing agreement
 - Confidentiality prohibits
 - Return all data upon completion
 - Send manuscript for disclosure limitation review
- Prepare for site visit

- All staff training in data confidentiality & security
- Data management
 - Immediate separation of identity
 - Only data manager can link data with identity
- embedded a hidden signature in each electronic file, for tracing unauthorized copies
- Site visits to monitor the use of these data at outside institutions

Sign agreement
- Comply with security requirement
- IRB approval plan for data security

Mixed mode sharing

Data enclave under supervision

Data archive for public use

Individuals and institutions are encouraged to develop and implement a similar training program.