Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance

World Health Organization

Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance

World Health Organization

Sharing and reuse of health-related data for research purposes:
WHO policy and implementation guidance

ISBN 978-92-4-004496-8 (electronic version)
ISBN 978-92-4-004497-5 (print version)

Suggested citation. Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance. Geneva: World Health Organization; 2022. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Cover photo credit: WHO/Hery Razafindralambo
Design and layout by ACW.

Contents

Foreword

When data related to research activities are shared ethically, equitably and efficiently, there are major gains for science and public health. However, the extent of sharing of health research data sets is lower than is needed to unlock these gains. This is why we are prioritizing the development of practical guidance to assist the many World Health Organization (WHO) colleagues who manage health research to enable greater sharing of research data. WHO staff members may be responsible for research projects funded or sponsored by WHO, or they may be involved in research in other ways – such as by coordinating networks of researchers or providing technical support to ministry of health staff who manage or fund research programmes.

As I write this Foreword we have seen the problems caused by the lack of data-sharing on COVID-19. Many thousands of COVID-19 clinical trials have taken place, but most have been too small or inadequately designed to provide useful evidence for policy-makers. A more mature ecosystem for curating and aggregating data sets would have been extremely helpful in collating the evidence base for decision-making. It is therefore very timely for WHO to publish this guidance which provides practical assistance to our staff in how to ensure that that final data from health research with which WHO is associated are further shared for reuse. WHO believes that the global scientific community should embrace the norm of sharing of data sets at the time of publication and, in some cases (such as pathogen genomic data), even prior to publication.

In parallel, the legitimate equity and career development needs of researchers around the world need to be addressed through initiatives that ensure access to medical products that are developed and to capacity development, including in data management and analysis. WHO's view is that the sharing of health data is a global public good. Countries that share data should be applauded, and never penalized, for contributing to that public good.

Dr Soumya Swaminathan
WHO Chief Scientist

Acknowledgements

This guide for WHO staff was developed by a WHO Science Division working group chaired by Vasee Moorthy, Senior Advisor, Research for Health.

Robert Terry was the lead writer. Main technical contributors were: Lisa Askie, Draurio Barreira Cravo Neto, Craig Burgess, Sarah Charnaud, Ian Coltart, Janet Diaz, Nathan Ford, Lisa Haintz-Carbonin, Ghassan Karam, Katherine Littler, Fuad Mirzayev, John Reeder, Andreas Reis, David Schellenberg, Anthony Solomon, Soe Soe Thwin and Sachiyo Yoshida.

WHO thank the following external expert reviewers: Philippe Guerin, Georgina Humphreys, Steve Kern, Rebecca Li, Rebecca Lawrence, Laura Merson, Sally Rumsey, and Tim Smith.

World Health Organization **policy** on the sharing and reuse of health-related data for research purposes

Vision

Great advances in science and public health can be achieved through the appropriate sharing and reuse of health data, permitting analyses that: 1) allow for the fullest possible understanding of health challenges; 2) help develop new solutions; and 3) ensure that decisions are based on the best available evidence. Data, and the knowledge derived from the use of that data, should be recognized as a global public good, and data-sharing and data reuse should be maximized in ways that are effective, ethical and equitable in order to improve public health.

Purpose and scope

The purpose of this document is to clarify the policy and practice on the reuse and onward sharing for research purposes of health data collected under the auspices of WHO technical programmes. This covers use in both emergency and non-emergency situations and complements the following: the *Policy on use and sharing of data collected in Member States by the World Health Organization (WHO) outside the context of public health emergencies*; the *Policy statement on data sharing by the World Health Organization in the context of public health emergencies*; and the *Joint statement on public disclosure of results from clinical trials*. This policy covers the reuse of health data for research purposes. Its scope includes research data generated by research undertaken directly by WHO, or funded by WHO, as well as the use of other health data for research purposes.

This document sets out the objectives of this WHO policy and the obligations of WHO staff and researchers funded by WHO. The following section of this document entitled *Guidance on the implementation of the WHO policy on the sharing and reuse of health-related data for research purposes* provides further references and resources to assist in the development of a data management and sharing plan that is in alignment with the vision of this policy.

Objectives for data-sharing for research purposes

The sharing of data collected under the auspices of WHO technical programmes must be undertaken in ways that are:

- Equitable – any approach to the sharing of data should recognize and balance the needs of: participants and researchers who generate and use data; other analysts who might want to reuse those data; and those communities who expect health benefits to arise from research.
- Ethical – all data sharing should balance and protect the privacy of individuals and the dignity of communities while acknowledging the imperative to improve public health through the most productive use of data.
- Efficient – any approach to data sharing should be aimed at enhancing/optimizing the quality and value of the use of those data and enabling their contribution to improving public health. Data sharing should be done as promptly and in as open a manner as possible, building on existing norms, policies and practices and reducing unnecessary duplication and competition.
- FAIR – WHO encourages data management and sharing to follow the FAIR principles (Box 1).

World Health Organization **policy** on the sharing and reuse of health-related data for research purposes

In summary, following these principles should ensure that the data are:

- Findable – manual or machine searchers can easily locate data using the Internet and a unique and persistent identifier.
- Accessible – once users find the required data, they need to know how those data can be accessed, possibly including authentication and authorization.
- Interoperable – which means that, where possible, the data are stored in the simplest nonproprietary software format (e.g. spreadsheet data stored as a CSV file rather than in a commercial software such as Excel).
- Reusable – in addition to the above, all other information or software required to access and use the data are provided, as are rich metadata (summary) describing clearly what the data contain and how they are organized under a clear and accessible data usage licence.

In order to address the above principles, WHO staff and researchers supported by WHO should develop a data management and sharing plan for each data set for which the Organization has responsibility in order to ensure the following:

- Data sets that contain no participant information, or which have been anonymized, should be deposited in an appropriate data repository with a persistent identifier, such as a Digital Object Identifier (DOI), and made available under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence, which clearly describes the terms of reuse.
- These data sets, described by accurate metadata, must include a description with links to the underlying data, individual participant data or extended data, and to any relevant materials or software necessary to understand, assess and replicate the research.
- The same conditions apply in the case of research articles that are funded in whole or in part by WHO. These must include a Data Availability Statement with links to underlying data or extended data and any relevant materials or software necessary to understand, assess and replicate the research.
- In cases where data cannot be made publicly available for ethical, legal and/or confidentiality reasons, a metadata record should be created in an appropriate data repository with a persistent identifier, such as a DOI. The Data Availability Statement should indicate the restrictions, the process for applying for access to the data, and the conditions that will apply for reuse.
- The chosen repository, hosted either by WHO or by a third party, should be able to demonstrate in a transparent manner how its governance arrangements meet any technical, ethical and legal requirements that may be applicable to it, including any national requirements as appropriate.
- Chosen repositories should maintain appropriate user agreements that govern the sharing and contribution of data stored by them. Data use agreements should address needs for sharing and reuse in line with this policy.

A data management and sharing plan must be developed as part of any technical programme or research project in order to guide the data collection, curation, storage, access, analysis, archiving and, in rare cases, disposal throughout the project or research cycle.

Guidance for **implementation** of the WHO policy on the sharing and reuse of health-related data for research purposes

Who is this guidance for?

The guidance is intended for WHO staff and other researchers who are funded in whole or in part by WHO.

Aim of the guidance document

The aim of this document is to provide guidance on how to develop a data management and sharing plan in order to share a WHO data set in a digital format in a way that is in line with WHO's policy on the sharing and reuse of health-related data for research purposes (the "Policy" or the "WHO Policy").

What data are covered?

The Policy covers both emergency and non-emergency situations and complements the following from the perspective of reuse: the *Policy statement on data sharing by the World Health Organization in the context of public health emergencies*; *(1)* the *Policy on use and sharing of data collected in Member States by the World Health Organization (WHO) outside the context of public health emergencies*; *(2)* and the *Joint statement on public disclosure of results from clinical trials*. *(3)* The scope of the Policy includes research data generated by research undertaken directly by WHO or funded by WHO, as well as the use of other health data for research purposes.

This guide is a **summary** of advice, providing references and links to more detailed resources. In many areas, there are no best practice recommendations or accepted standards. In such cases, the practices that WHO considers to be good practice are set out in the Policy.[1] Consequently, users will need to exercise judgement, taking into account all relevant considerations, in order to identify the most appropriate mechanism for storing and sharing specific data set/s. For instance, sharing pathogen genomic data requires different considerations than sharing individual patient data (IPD) that underlies a clinical trial.[2]

It is recommended that users of this guide develop **a data management and sharing plan** (see Annex 1) in line with the principles and objectives of the Policy. The plan should meet the WHO principles of being effective, ethical and equitable – as articulated in the Policy. The plan should also follow the technical FAIR principles as far as possible and should aim to make the data as open for reuse as possible while balancing the necessary ethical and legal factors that protect the interests of the stakeholders related to the data set.

1 This guide may contain links to third party resources or external websites. WHO is not responsible for the accuracy or content of any third party resources or external links. Reference to any third party resource or external link does not imply that the resource or link, or their author or entity, is endorsed or recommended by WHO. These references are provided for convenience only.

2 in this document we use IPD to mean individual participant data to cover all instances of WHO research that involve people and their personal data. Please note in clinical research IPD is more typically defined as individual patient data.

BOX 1. The FAIR Guiding Principles

To be findable:

F1 (meta)data are assigned a globally unique and persistent identifier
F2 data are described with rich metadata (defined by R1 below)
F3 metadata clearly and explicitly include the identifier to the data it describes
F4 (meta)data are registered or indexed in a searchable resource.

To be accessible:

A1.0 (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2.0 metadata are accessible, even when the data are no longer available.

To be interoperable:

I1 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2 (meta)data use vocabularies that follow FAIR principles
I3 (meta)data include qualified references to other (meta)data.

To be reusable:

R1.0 (meta)data are richly described with a plurality of accurate and relevant attributes
R1.1 (meta)data are released with a clear and accessible data usage license
R1.2 (meta)data are associated with detailed provenance
R1.3 (meta)data meet domain-relevant community standards.

This approach usually means choosing an appropriate third party repository and curating the data so that they are: 1) well described using a clear summary (metadata); 2) structured so they can be reused and aggregated with similar data sets; and 3) where necessary, anonymized to reduce the possibility of individual participants being identified. The chosen repository must be assessed to ensure that it meets the principles outlined in this document – particularly where the data are sensitive (e.g. in the case of IPD). There are some instances where WHO may directly manage a repository within its own resources; WHO-funded researchers should investigate whether their university or host institute has established its own data repository or set up an institutional subscription with a data-sharing organization.

How to use this guide

The recommendation is that a data management and sharing plan should be developed before any data collection has begun. Nevertheless, there are still many practical solutions to sharing data in an ethical and effective way, even after a programme of activity has completed. WHO encourages users to read the various sections of this guide and follow up on those resources that are most relevant to their situation. It is recommended that the template in Annex 1 be used to draft a data management and sharing plan. WHO-funded researchers may choose to discuss this with the WHO technical programme that supported their work. WHO will undertake periodic reviews of these plans in order to further improve this guidance, particularly where standards and best practices are developed.

Where can I get help and advice?

Help and advice are available through the Science Division. Please contact Rob Terry for technical advice, Ian Coltart for licensing/publication requests, and Vasee Moorthy for questions on genomic data sharing. Where it is unclear who to contact, these colleagues can triage requests to the correct staff member.

Making a data management and sharing plan

Why?

Any proposal to share data requires careful consideration of the relevant technical, legal and ethical issues. Users need to understand the data that are available and how they can be used. To this end, a structured description of the data set, known as the metadata, as well as a Data Availability Statement, are needed. The benefits of developing a data-sharing plan include the following:

- A data-sharing plan demonstrates good research practice and marks the completion of the research cycle.
- Data sets can be used and cited by others, providing mechanisms of credit for the work involved in creating the data set.[3]
- Data sets can be aggregated into other studies, can be reproduced and can enable the creation of new knowledge for the benefit of science and public health.
- The plan ensures appropriate protection and privacy of participants where relevant.
- The plan ensures preservation and long-term archiving of the data.
- A data-sharing plan satisfies the requirements of many journals and funders, including WHO.

The type of data will influence the data-sharing plan required

Understanding the data set will help decide how data sharing can be achieved. Many universities provide advice through their library services, and a number of journals and data organizations have published guidance. The simplest entry guides on understanding the sharing of research data have recently been published by the F1000Research publisher (2021). This describes what open data is, how and where to choose a repository, how to structure data (e.g. in spreadsheets), applying the FAIR principles, and how to address legal and ethical issues such as achieving anonymization. *(4)*

3 A 2019 study estimated an increase of 25% in citations where data were made open. See: Fane B, Ayris P, Hahnel M, Hrynaszkiewicz I, Baynes G, Farrell E. The state of open data report 2019. Digital Science. 2019. https://doi.org/10.6084/m9.figshare.9980783.v2.

A comprehensive resource for managing clinical health data for research purposes has recently been published by the Global Health Network that created a knowledge hub for the European and Developing Countries Clinical Trials Partnership (EDCTP) which was launched in 2021. *(5)* This resource sets out the knowledge management steps throughout the research cycle, explains how to develop a data management and sharing plan, and provides a repository finder tool and other resources. In addition, the Global Health Network offers free online **training courses** in data management and sharing, as well as training on general research skills – with certification upon completion. *(6,7)* These courses have been designed to suit clinical research and researchers in low- and middle-income settings. The knowledge hub also has links to many other resources. *(8)*

What information does the data set describe?

The main concerns when planning to share health data relate to the history and provenance of the data. They include:

- Do the data contain any information that could directly identify an individual (e.g. name, telephone number, email) or indirectly identify a person by linking a detailed geographical location with a disease?
- Has participant or patient consent been given for reuse and what specifically does the consent allow in terms of reuse?

Where data are collected under a research protocol (e.g. clinical trial) the consent obtained should allow for publication of the analysis and deposition in a repository of the underlying data reported in a study. This type of consent will **not** cover the sharing of individual participant data unless it has been adequately anonymized (see below). One prerequisite for further sharing of data is that the data-sharing agreement includes appropriate legal provisions governing receipt of data by WHO or the WHO-funded researcher.

For WHO, any processing of personal data[4] should be done in compliance with WHO's Personal Data Protection Policy. The "Research Data" section of the Personal Data Protection Policy applies in cases where the direct purpose of the processing of personal data is scientific research.

Sharing identifiable participant (or patient) data (IPD) requires more planning but can be safely achieved to answer extremely important health questions to which answers may not be possible from single studies. An example would be to identify groups that are under-represented in a number of research studies in order to understand, for example, contraindications in medicines with respect to children or pregnant women. This may also include analysing electronic patient records using data collected as part of service delivery but that do not have a standing consent, provided that the data provider has secured all necessary permissions according to the applicable national laws of the country where the data were collected. Several solutions are available to enable sharing through managed access repositories. Further information is provided below.

Users also need to understand what formats their data are in – for instance,

4 Personal data: Any information relating to an individual who is or can be identified from that information. Personal data include: biographical data (biodata) such as name, sex, civil status, date and place of birth, country of origin, country of residence, individual registration number, occupation, religion and ethnicity; biometric data such as a photograph, fingerprint, facial or iris image; as well as any expression of opinion about the individual, such as assessments of status and/or specific needs. Processing of personal data: Any operation, or set of operations, automated or not, which is performed on personal data, including but not limited to the data collection, recording, organization, structuring, storage, adaption or alteration, retrieval, consultation, use, transfer (whether in computerized, oral or written form), dissemination or otherwise making available, correction or destruction.

images (e.g. MRI scans), audiovisual formats (e.g. interviews), Office software (e.g. Excel), computer software programmes, a database (e.g. REDCap), DNA sequence information etc. Many different repositories exist to handle these types of data and it is likely that, with some research, a suitable resource will be found.

Which data can be shared openly?

In most cases, data that contain no participant information (e.g. a survey of insect vectors) or that have been anonymized can be shared by deposition in an appropriate repository that complies with the WHO Policy. This guide will help you decide which repository is best suited to your data.

Which licence can be used to share data?

The WHO Policy requires that WHO's data are shared under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. This open licence means users must acknowledge the use of the data in any subsequent publication. The data set must be cited – e.g. using the DOI assigned to the data set when deposited. However, this type of licence means that those responsible for sharing the data do not need to be contacted or involved in the reuse. This is analogous to someone citing a research paper authored by someone else. This makes the data sharing more efficient and increases its reusability. The option of applying this licence is usually available when depositing the data.

Preparing the data for sharing

The data-sharing toolkit of the EDCTP Knowledge Hub describes the major steps in preparing data for sharing. *(9)* These steps have been adapted and are summarized below. They should form the basis of the Data Management and Sharing Plan. The EDCTP toolkit provides further detailed advice for each step.

1. Choose a suitable repository and set up an account.
2. Understand the terms and conditions of the Data Deposit Licence. In cases where WHO wishes to conclude an agreement with a third party repository, the agreement must be reviewed and cleared by the WHO Legal Office. WHO-funded researchers are responsible for the data they collect and for concluding their own agreement with the repository.
3. Clarify ownership, rights and permissions. Ensure that rights have been obtained to share the data – and, if appropriate, that existing consent covers the data sharing and reuse intended – and define any conditions that apply. Ensure that the Data Deposit Licence does not conflict with WHO Policy by placing additional restrictions on further reuse.
4. Structure and organize the data in files that are logical and clearly named.
5. Check that the data are accurate and consistently labelled. Some disciplines, particularly in chemistry and the omics, have a recommended structure for reporting and depositing in a repository. Consider also supplying non-proprietary files – e.g. CSV files as well as structured data sheets such as Excel. This maximizes reuse (i.e. users do not require MS Office to use the files) and long-term preservation (i.e. when proprietary file formats become obsolete). *(10)*
6. Prepare supporting documents, user guides and any tools (e.g. software) needed to understand and use the data and deposit them along with the data. This may require that a data dictionary is developed to define each of the variables in the data set.
7. Deposit the data and obtain a DOI that can be used to cite the data set in any reports or journal articles as part of the Data Availability Statement.

Some of these steps are expanded below with specific reference to the requirements of the WHO Policy.

Please note: This is only a short and introductory summary. Data depositors should ensure that they have access to the required expertise in making a final decision.

Choosing a data repository

Table 1 provides a simple decision tree as a guide to choosing a suitable data repository. The links give access to online search tools. However, the list is not exhaustive, so it is advised to consult the resources available in the list of references at the end of this document in order to make an informed choice. Where specific platforms are referenced, these are provided for information and should not be considered an endorsement or recommendation by WHO. Researchers funded by WHO should check if their university or host institution has its own data repository or an account with a general data repository that is compliant with WHO's policy. Most repositories will require registration. Some repositories offer a free service up to a specified data limit and then a subscription service beyond that; others charge for all deposits. Where a charge to deposit applies, it is worth investigating if there are waivers or discounts for researchers from low- and middle-income countries (LMICs).

What if it is not possible to share the data set openly?

In certain cases it may not be possible to deposit the data set in a repository. This might be for legal reasons (e.g. where some of the data were obtained from a proprietary source) or for ethical reasons when consent or appropriate permission cannot be obtained for sharing data (e.g. electronic records) from a Member State.

In these cases, two options are available, although neither is ideal and they should be considered only as a last resort. One option is to retain the data within WHO or at the institution of the WHO-funded researcher's institute. If a publication is made with these data, a Data Availability Statement should be included with the publication.

A second option is possible where it is not planned to use the data in a publication. In this case, the data depositor should prepare a description of the data set – known as the metadata – and deposit that description (but not any related data files) in a repository. This will also generate a specific web address (DOI) that can be used to cite the summary of the data easily. Within the metadata the depositor can also include instructions on how others can make contact and discuss if there is a possibility to access or use those data. This is analogous to the Data Availability Statement in a publication. It will be desirable to have the necessary procedures in place to handle a request to access and use these sensitive data sets, including a sound rationale for refusal if necessary. There remains an obligation to ensure that the data – e.g. clinical trial data – are stored and archived properly for future reference. While not ideal, the creation of a summary at least makes available an inventory of WHO data. Please seek advice if you are unsure how to proceed.

Consent and data sharing for health-related purposes.

There is no single approach to obtaining consent for research purposes. WHO has worked with the Council for International Organizations of Medical Sciences (CIOMS) to produce ethical guidelines for health-related research involving humans. *(11)* These guidelines are the basis for creating sound consent processes, and a number of templates are used by WHO to match different contexts. With

respect to data sharing, CIOMS Guideline 12 on Collection, storage and use of data in health-related research states that:

"Researchers, sponsors and research ethics committees must share data for further research where possible."

Guideline 12 also sets out the conditions where a research ethics committee can grant a waiver for informed consent on the sharing of stored data where no individual consent exists:

"When researchers seek to use stored data collected for past research, clinical or other purposes without having obtained informed consent for their future use for research, the research ethics committee may consider to waive the requirement of individual informed consent if:

1. the research would not be feasible or practicable to carry out without the waiver; and
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." *(12)*

Further conditions for data sharing are spelled out in Guideline 24 on Public accountability for health-related research. Guideline 24 states that:

"Researchers should prospectively register their studies, **publish the results and share the data on which these results are based in a timely manner**. Negative and inconclusive as well as positive results of all studies should be published or otherwise be made publicly available."

There is a following commentary on this guideline which states:

"There are compelling reasons to share the data of health-related research. Responsible sharing of clinical trial data serves the public interest by strengthening the science that is the foundation of safe and effective clinical care and public health practice. Sharing also fosters sound regulatory decisions, generates new research hypotheses, and increases the scientific knowledge gained from the contributions of clinical trial participants, the efforts of clinical trial researchers, and the resources of clinical trial funders …"

"… 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**." *(13)*

In summary, these guidelines support and encourage data sharing when the appropriate governance processes are in place to protect individual privacy. The CIOMS guidelines underpin many Good Clinical Practice documents, including the guidance provided by WHO. *(14)*

In drafting a consent form that adequately covers the secondary use of the data, *(15)* the United Kingdom's Data Service has produced useful guidance, particularly with respect to language to be avoided. It advises that:

"Consent forms should not preclude sharing of research data. So, promises to destroy any data or that data will only be seen or accessed by the research team should be avoided."

"Terms such as 'fully anonymous' or 'strictly confidential' should be avoided, as they are often impossible to define. Better is to indicate how data will be anonymized (e.g. by removing all personal information that could directly identify an individual) and that whilst data will be made available to other researchers, confidentiality will be protected." *(16)*

Anonymization and de-identification

Data variables are categorized as direct and indirect identifiers. Direct identifiers link directly to a person's identity without further information (e.g. name, address, medical record number, photograph, etc). Indirect identifiers require some deduction to identify the person and include, for instance, a higher level of address, disease condition, gender, generic job title, date of birth. By combining these variables, there is a reasonable probability that an individual can be identified, (e.g. a 50-year-old man with diabetes on a certain street in Geneva).

"Anonymization" is the process by which those variables that can identify an individual are irreversibly removed from a data set. "De-identification" is the process by which these identifiers are masked in some way or replaced by a code that has a key. Applying the key can reinsert the identifiers to enable re-identification. As such, control of the key is limited and its use allowed only in certain circumstances. De-identification through an encrypted code and key is a common approach used in clinical trials.

There are no global standards for anonymization or de-identification. Many countries have adopted data protection laws describing what constitutes anonymous data. It is important to be aware of any applicable legal frameworks, including in the place of collection and storage of the data. *(17)*

Many data managers within the research community use the standards recommended in the United States of America or in Europe, depending on their respective places of business and applicable legal frameworks. *(18, 19)*

In addition, the Personal Data Protection Commission of Singapore and the Australian National Data Service (ANDS) have both produced useful guides. *(20, 21)* An infographic from ANDS is reproduced in Annex 2. Calculation of the risk of identification is itself an imprecise measure and there are alternative ways to measure it. *(22)*

WHO-funded researchers should seek advice from within their university or host institution and should contact WHO if they require further assistance.

Choosing an appropriate repository

Repositories can be organized in a number of ways. The simplest repositories are often free to use up to a certain data limit with deposition and access via a website without any mediation or review by committee. Consequently, they can be described as open repositories. There is a structure for describing the data with a summary – the metadata – but there is no requirement to format the data to a certain standard. Open repositories are sometimes called data lakes, as the data are all in one place but in variety of formats with different standards. Reuse conditions are set by the depositor from "open", under a Creative Commons licence, to "closed" whereby only the metadata are available. Examples of open repositories include Figshare, Mendeley and Zenodo.

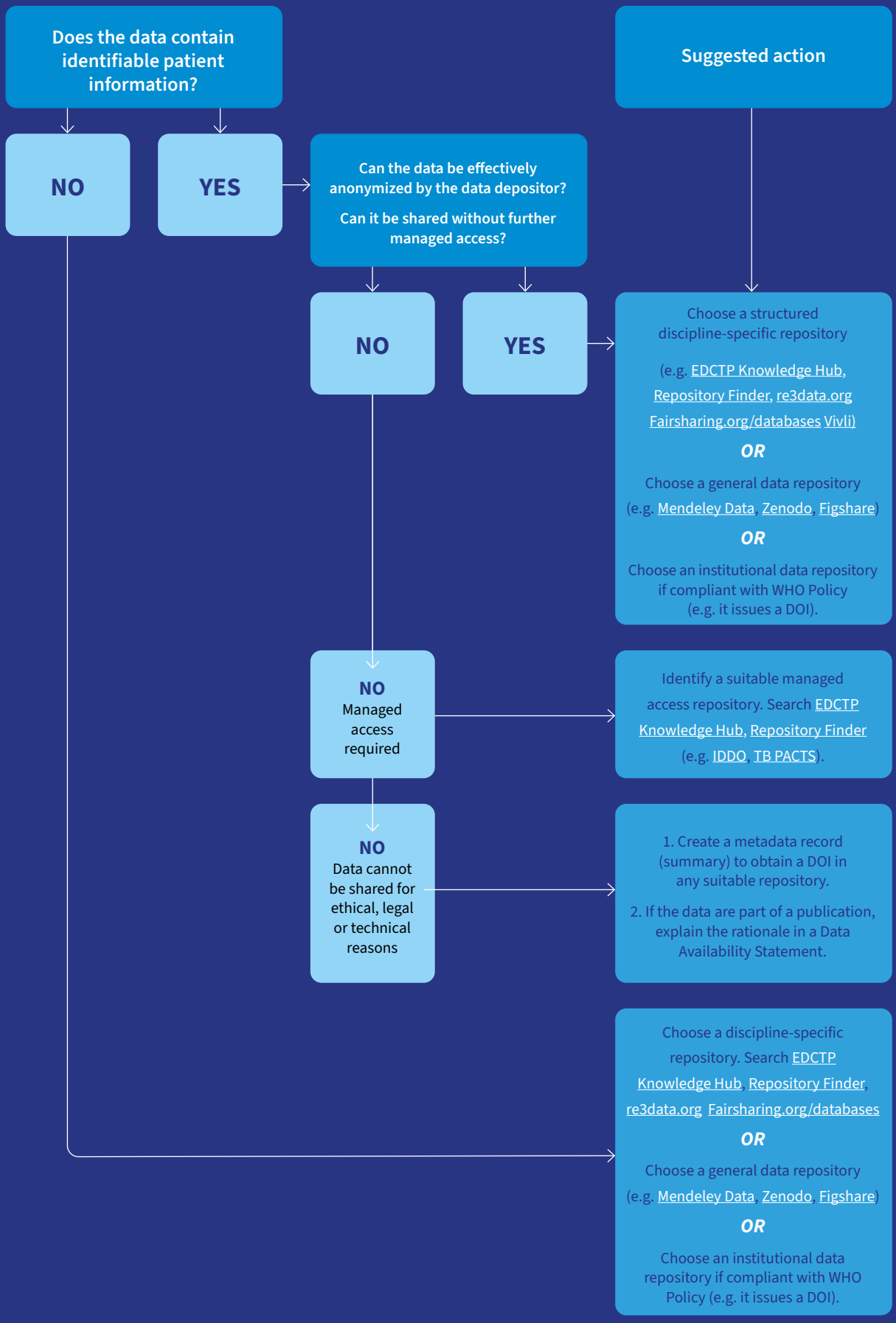
Other repositories may include more structure for deposition, with certain conditions for access – such as a requirement to structure the data according to a certain standard or a restriction on the republishing of those data elsewhere. For some repositories, particularly those dealing with more complex and sensitive data such as IPD, the steps for depositing and accessing the data are moderated. These managed resources may include reviews by data access committees, data transfer agreements and some continued involvement of the depositor in the type of reuse. There may be a requirement to deposit data in a certain standard structure or, alternatively, curation to a standard structure may be provided by the repository. There are also different ways in which the data may be stored, either together in a centralized location, or kept in different places but accessed through a specialized website – sometimes known as a federated system. There are numerous variations of these repositories, each with its own benefits and drawbacks.

If the data being shared are relatively simple and there are no concerns about privacy the open repositories offer a straightforward solution. The data are securely archived, free to access and the depositor has no need to be involved in further research. The route to access and reuse of the data is quick and simple. However, as there are no requirements for specific standards, persons accessing the data will have to undertake their own curation when combining data from different sources.

Managed platforms require more resources and the process of accessing the data may be lengthy, requiring review from a data access committee and the resolution of questions regarding the protocol. However, if numerous data from many sources (e.g. electronic patient records from many different hospitals in different countries) are combined, the fact that the data are in a single standard (e.g. C-DISC for participant/patient data) means that the analysis can proceed more quickly and the utility of the whole data set is greatly enhanced.

The simple decision tree in Table 1 can guide users into making the first level of decision about where to deposit their data. Again, if you require assistance in making this choice, please contact: terryr@who.int, coltarti@who.int or moorthyv@who.int .

TABLE 1. Decision table – how to choose the right repository



Clinical trials

A number of established good practice arrangements exist for clinical trials and their reporting of the underlying data. WHO has issued a Joint statement on public disclosure of results from clinical trials.3 The statement covers the registration of a trial prior to commencement, the use of the trial ID in all subsequent communications, the timely reporting of results within 12 months from primary study completion (the last visit of the last subject for collection of data on the primary outcome) in a trial registry and no longer than 24 months after trial completion in a journal publication.

Sharing individual participant data (IPD) in research

IPD can be shared in an appropriate public repository, as listed in the repository finder tools highlighted in Table 1, provided that the data are adequately anonymized. These data sets can be used in systematic reviews to aggregate reported results into a finding with a greater degree of confidence. As noted above, any processing of personal data by WHO should be done in compliance with the WHO Personal Data Protection Policy.

Equity in data sharing and capacity-building

A fundamental objective of the WHO Policy is that data are shared in ways that are ethical, efficient and equitable. For WHO, equitable data sharing means balancing the needs of all stakeholders – the data depositors, those generating the data (including researchers in the case of research data sets), the secondary users of data sets, and the communities from where the data have originated. WHO considers it to be crucial to address the needs of data depositors and explores every avenue to advance equity and capacity-building, including within the area of data sharing.

With regard to research data sets, many researchers in LMICs consider that data sharing policies – particularly those of international funders of research – may disadvantage them if the WHO Policy requires data to be shared within short timescales or in ways that do not involve the researchers in the analysis. WHO will promote data agreements – for instance, through advocacy with funders of international research –that include involvement of researchers from the countries where data are collected. All subsequent publications should reflect those contributions in the authorship.

Capacity-building is closely linked to achieving equitable approaches to data sharing. WHO seeks to build research capacity through its technical programmes that support research and training courses in LMICs, as well as other support. This includes efforts to provide training in data management and curation as well as the skills required to manage and analyse health-related data.

Which Creative Commons licence should be used for data that are made openly available?

The data should be shared under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. This means the data are free to use without seeking further permission from the depositor but the users must provide a proper attribution or citation as to where they obtained the data from. This is why obtaining a DOI is useful as it provides a precise and permanent location for the data set so that the depositor can receive citation credit.

Describing the data set with metadata

Metadata is "data about the data" which describe the properties of the data set. Metadata should be structured so that it can be machine-read (e.g. by an Internet search engine) to maximize findability. *(23)* The elements covered by a metadata statement are often prescribed by the repository used and commonly include the following:

- **Title**: This is often the title of the research project or, if the data are from a technical programme, a description of that programme and the type of data included.
- **Owner**: A contact email and an address should be provided for the data creator analogous to the corresponding author on a research paper. If the data belong to WHO, then WHO might be added as the publisher of the data.
- **Contributors:** Analogous to co-authors in research manuscripts, the term "contributor" describes who else should be credited with creating the data set.
- **Subject**: These are keywords and some repositories will provide drop-down menus (e.g. MeSH). The subject may include when, where and how the data were collected and the type of data (e.g. trial data, household survey).
- **Description:** As in an abstract on a research paper, the is a brief description of the data. This may include information describing the structure and titles of the data files.
- **Licence**: The WHO Policy requires Creative Commons Attribution 4.0 International (CC BY 4.0).
- **Date**: The date when the metadata were created, plus any modifications. This is often automatically added by the repository.
- **Format:** The standard by which a file is encoded. A free standard is preferred. However, if a proprietary standard is used, copies should be supplied in non-proprietary files where possible (e.g. alongside common files such as REDCap or Excel).
- **Related publication**: A url or DOI of any related publications and journal articles should be included.
- **Language**: The language(s) in which the data are presented.

Writing a Data Availability Statement

This is increasingly required as part of the submission process for publishing in a journal and is recommended by the International Committee of Medical Journal Editors (ICMJE). If the data set is deposited in a repository, the following headings should be included:

- repository name
- title of the data set
- DOI
- list of all data items (including the full file name, and a description of its contents)
- data licence
- any restrictions on access to the data.

N.B. Where data are NOT shared through a repository, a statement to this effect is required and should include details of who may be contacted – and how – by interested parties wishing to discuss access to the data.

References

1 Policy statement on data sharing by the World Health Organization in the context of public health emergencies. Geneva: World Health Organization; 2016 (https://www.who.int/ihr/procedures/SPG_data_sharing.pdf, accessed 29 January 2022).

2 Policy on use and sharing of data collected in Member States by the World Health Organization (WHO) outside the context of public health emergencies. Geneva: World Health Organization; 2017 (https://www.who.int/publishing/datapolicy/Policy_data_sharing_non_emergency_final.pdf, accessed 29 January 2022).

3 Joint statement on public disclosure of results from clinical trials. Geneva: World Health Organization; 2017 (https://www.who.int/news/item/18-05-2017-joint-statement-on-registration, accessed 29 January 2022).

4 Understanding open data. London: F1000 Research Ltd; 2021 (https://think.f1000research.com/open-data/, accessed 28 January 2022).

5 EDCTP Knowledge Hub. European and Developing Countries Clinical Trials Partnership Data management portal (https://edctp-knowledgehub.tghn.org/Dat-man-por/, accessed 28 January 2022).

6 EDCTP e-Learning courses. European and Developing Countries Clinical Trials Partnership Data management portal (https://edctp-knowledgehub.tghn.org/training-courses/e-learning-courses/, accessed 28 January 2022).

7 Global Health Training Centre (https://globalhealthtrainingcentre.tghn.org/, accessed 28 January 2022).

8 Free tools and resources. European and Developing Countries Clinical Trials Partnership Data management portal (https://edctp-knowledgehub.tghn.org/data-sharing-toolkit/collated-external-resources/, accessed 28 January 2022).

9 EDCTP Data sharing steps. European and Developing Countries Clinical Trials Partnership Data management portal (https://edctp-knowledgehub.tghn.org/data-sharing-toolkit/data-sharing/, accessed 29 January 2022).

10 EDCTP Formats list. European and Developing Countries Clinical Trials Partnership Data management portal (https://edctpknowledgehub.tghn.org/data-sharing-toolkit/data-sharing/formats-list/, accessed 29 January 2022).

11 International Ethical Guidelines for Health-related Research Involving Humans, fourth edition. Geneva: Council for International Organizations of Medical Sciences (CIOMS); 2016 (https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf , accessed 29 January 2022).

12 International Ethical Guidelines for Health-related Research Involving Humans, fourth edition, Guideline 12. Geneva: Council for International Organizations of Medical Sciences (CIOMS); 2016:47–52 (https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf, accessed 29 January 2022).

13 International Ethical Guidelines for Health-related Research Involving Humans, fourth edition, Guideline 24. Geneva: Council for International Organizations of Medical Sciences (CIOMS); 2016:91–93 (https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf, accessed 29 January 2022).

14 Handbook for good clinical research practice (GCP): guidance for implementation. Geneva: World Health Organization; 2005 (https://apps.who.int/iris/handle/10665/43392, accessed 29 January 2022).

15 Research data management. Colchester: United Kingdom Data Service, University of Essex (https://www.ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing.aspx, accessed 29 January 2022).

16 Documenting consent. Colchester: United Kingdom Data Service, University of Essex (https://ukdataservice.ac.uk/app/uploads/documenting-consent.pdf, accessed 29 January 2022).

17 Greenleaf G. Global tables of data privacy laws and bills, fifth edition. 145 Privacy Laws & Business International Report. 2017;14–26 (https://ssrn.com/abstract=2992986, accessed 29 January 2022).

18 Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Washington (DC): US Department of Health & Human Services (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard, accessed 29 January 2022).

19 The European Medicines Agency (EMA) approach to anonymization draws most of its reference from the US Institute of Medicine report: Sharing clinical trial data: maximizing benefits, minimizing risk. Washington (DC): National Academies Press; 2015 (https://pubmed.ncbi.nlm.nih.gov/25590113/, accessed 29 January 2022).

20 Guide to basic data anonymisation techniques. Singapore: Personal Data Protection Commission; 2018 (https://iapp.org/media/pdf/resource_center/Guide_to_Anonymisation.pdf , accessed 29 January 2022).

21 De-identifying your data. Canberra: Australian National Data Service; 2018 (http://www.ands.org.au/working-with-data/sensitive-data/de-identifying-data, accessed 29 January 2022).

22 El Emam K, Abdallah K. De-identifying clinical trials data. Cranbury (NJ): Applied Clinical Trials; 2015 (https://www.appliedclinicaltrialsonline.com/view/de-identifying-clinical-trials-data, accessed 29 January 2022).

23 Describe: metadata and documentation. European and Developing Countries Clinical Trials Partnership Data management portal (https://edctpknowledgehub.tghn.org/data-sharing-toolkit/data-management-sub/Describe-Metadata/, accessed 31 January 2022).

ANNEX 1. DRAFT template for creating a data management and sharing plan

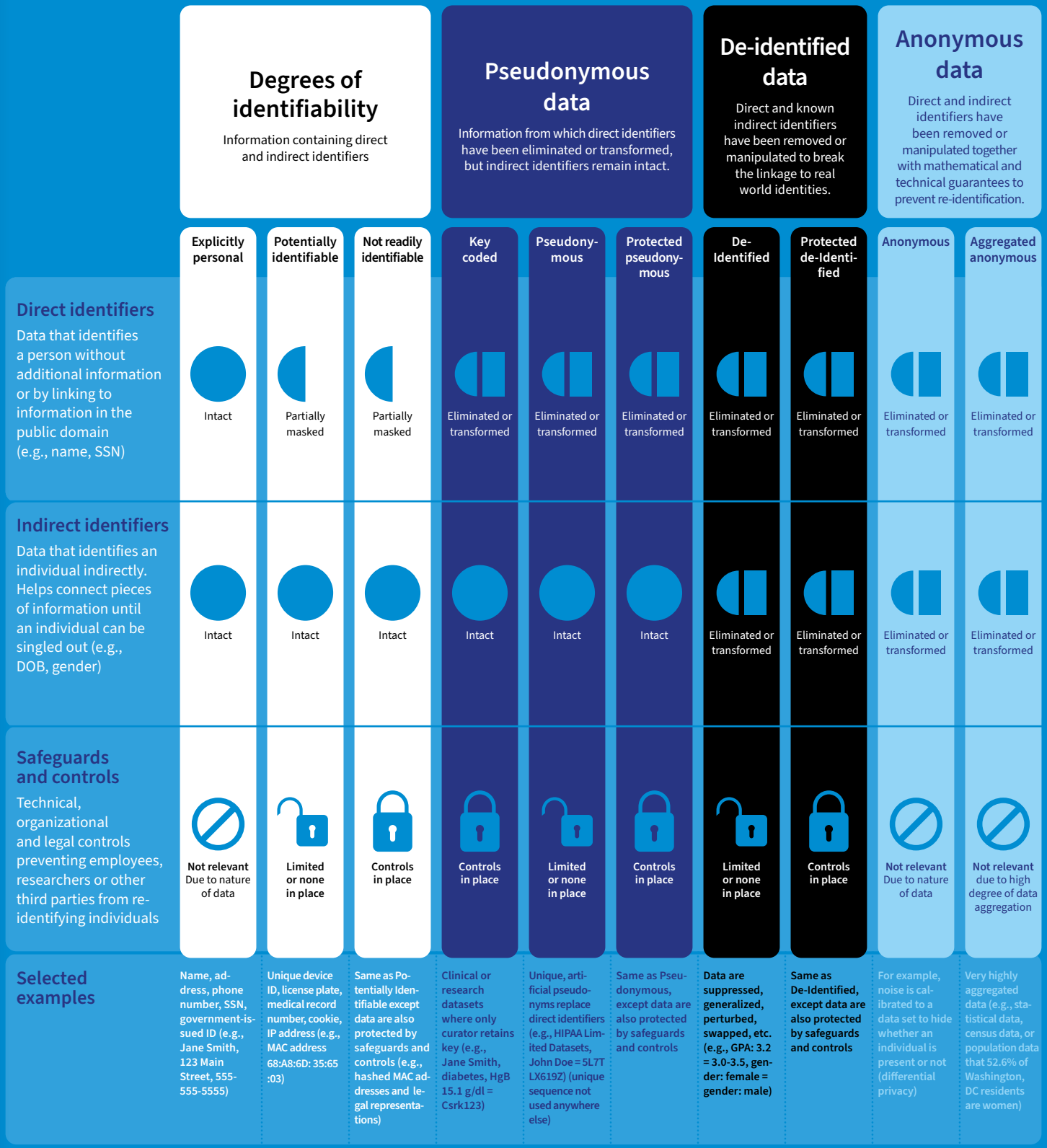
In preparing a data management and data sharing plan, WHO-funded researchers should seek advice from within their university or host institution. Use the template below as a checklist to develop your data management and sharing plan.

Topic	Action	Notes
What type of data?		
Who are the stakeholders?		
What rights and permissions – do you have? – do you need to obtain?		
Choosing a repository	Third party or WHO?	
Will you need to sign a data deposit agreement?		For WHO, review and clearance by WHO Legal office
Choosing NOT to share	Create a metadata record	Do you have a process in place to deal with enquiries for access?
Where appropriate, what consent exists? Do you need to apply for a waiver of requirement for consent?		
If necessary, what level of anonymization or de-identification do you need to apply?		
Clinical trials	Register	Decide when and how you will report
Sharing IPD: – open repository? – managed repository? – WHO repository?		
Equity		How can equity in data sharing, access and/or capacity-building be ensured?
Timelines for sharing		
Choosing the right licence	WHO Policy requires Creative Commons Attribution 4.0 International (CC BY 4.0) licence	
Create your metadata		
Write a Data Availability Statement		

Source: Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Nature: Sci Data. 2016;3(160018) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/pdf/sdata201618.pdf, accessed 31 January 2022).

ANNEX 2. Visual guide to practical data de-identification *(24)*

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability. **This is a primer on how to distinguish different categories of data.**



24 This infographic is adapted from the Australian National Data Service document a Visual-Guide-to-Practical-Data-DeID.pdf (2018) (CC-BY-ND 4.0) http://www.ands.org.au/working-with-data/sensitive-data/de-identifying-data.

Notes

World Health Organization
20, Avenue Appia
1211 Geneva 27
Switzerland

www.who.int

9789240044968