Ethical data governance for mental health databanks

A framework for risk diagnosis
and mitigation strategies

Contents

Click on the sections to navigate the document

Introduction

Wellcome's mental health strategy aims to create a step change in early intervention for anxiety, depression and psychosis. As part of this, Wellcome funds work that will transform understanding of how these conditions develop and resolve. They view longitudinal data as a key resource that can help achieve this goal.

Wellcome has commissioned a series of projects to explore global mental health *datasets* (a collection of data from a single source or for a single purpose) and *databanks* (large aggregations of data from many datasets) including:

- Identifying 'active ingredients'–the aspects of interventions that make them effective in preventing or managing anxiety and depression in young people
- The MindKind Study which explored how best to collect longitudinal youth mental health data
- Bridging the Gap which described technical, design and governance tools needed to scale participatory methods and support reproducible and inclusive science
- A landscaping report with an inventory of key mental health and related longitudinal datasets around the world

What we were commissioned to do

We add to this growing body of knowledge with this **critical ethical analysis framework** to support understanding and management of **ethical risks arising from the creation, enrichment, and aggregation of potentially sensitive datasets**. Although this work was commissioned as part of a mental health databanking effort, the framework highlights ethical considerations and mitigations that are more broadly applicable.

How we developed this framework

We developed this framework in three stages:

1. **Secondary research:** Reviewed existing literature on data governance for sensitive personal health data and its associated risks
- Landscaped existing literature (primary, grey) on mental health data governance
- Explored emerging ethical frameworks for mental health data
- Identified gaps in current literature and pathways for community engagement
- See *"Key novel and emerging ethical considerations"*

2. **Analysis:** Consolidated key considerations and mitigations for each of the risk areas we identified in order to iteratively develop the critical analysis framework
- Coalesced learnings from literature review
- Extracted mitigation strategies and synthesised outputs
- Conducted iterative refinement cycles to create this task-grounded ethical framework for databank builders
- Extracted in depth guidance (see *"Guide for individual dataset evaluation"*) to support databanks as they consider aggregating or supplementing a given dataset, including recommendations for selection criteria and documentation

3. **Sense checking:** Engaged lived experience advisors and subject matter experts to review the framework, ensuring that critical nuance was not lost through the synthesis process
- Hosted co-design workshops with Wellcome Trust's Lived Experience team for in-depth review and trial application of this framework
- Interviewed external experts and relevant Wellcome Trust stakeholders to check the framework for face validity (i.e., how well a tool, at the surface, represents what it is supposed to be measuring or describing)
- See *"Key Insights: Lived experience advisors and subject matter experts"*

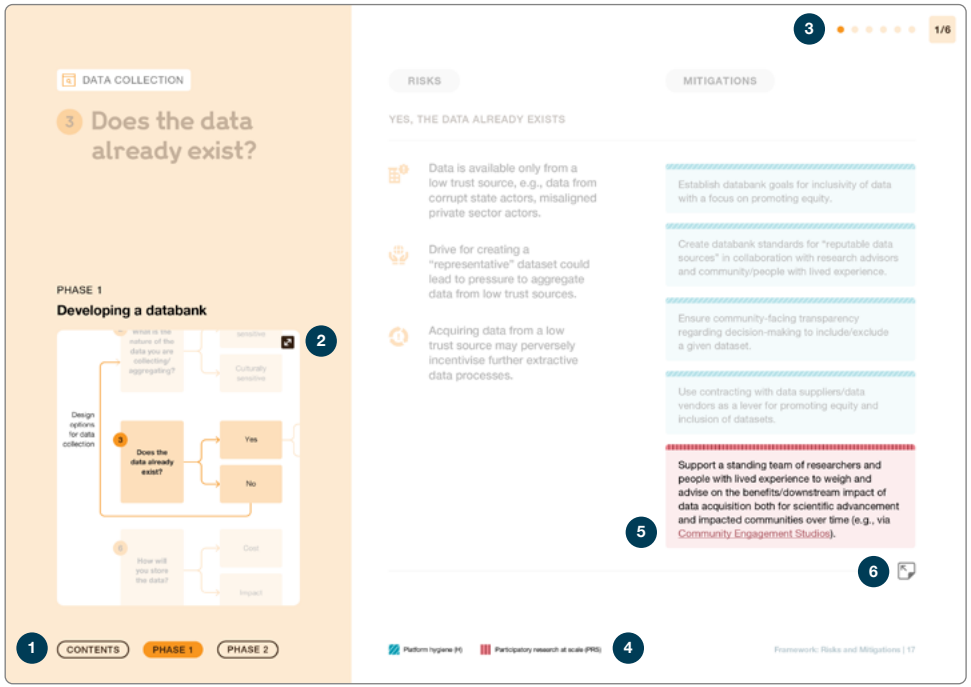
How you can use the framework

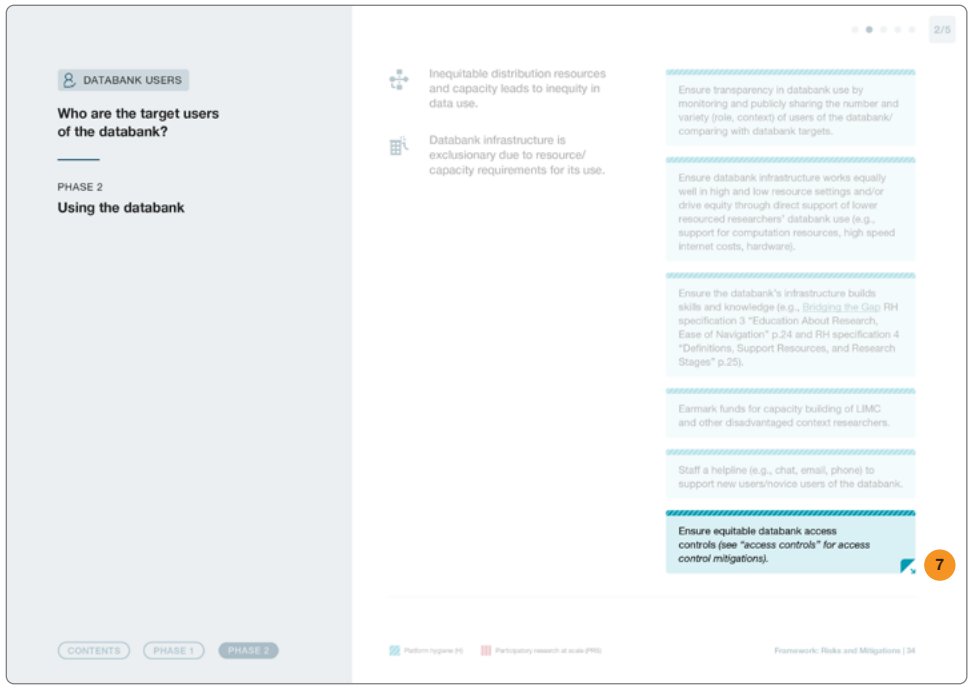
We developed this framework to guide databank builders through key ethical considerations associated with building a databank. We provide a library of mitigation strategies that databank builders can consider implementing as part of their work to address these considerations.

Others may also find this framework useful:

- Funders, governments, and civil society organisations/non-governmental organisations might use or build from this framework to guide evaluation of proposals for data collection, enrichment, aggregation, and/or use
- Research institutions and individual researchers could use this framework to extend the conceptualisation, planning, and execution of their work towards more equitable and community-centering outcomes
- Communities and people with lived experience are invited to use and adapt this framework to meet their needs as they evaluate opportunities for engaging with the research ecosystem

How to read this document





1 **Navigate easily:** Use the buttons at the bottom left corner of every page to access different sections of the document. Pages that contain other ways of navigating contain instructions marked with a ***cursor icon***

2 **Locate your position within each phase:** The snapshot on the left of each new section shows you where this section is located within the overall phase. To return to the larger flowchart, click on the ***expand icon***

3 **Locate your position within each section:** The progress bar at the top right corner of the page indicates the number of pages within that section and how much you have covered so far. Click to navigate within each section.

4 **Refer to the key:**
Mitigations related to platform hygiene are marked in ***teal with diagonal stripes***
Those related to participatory research at scale are marked in ***pink with vertical stripes***
Use the glossary to read further about these two categories.

5 **Read further:** Look out for underlined text. These have been hyperlinked to further information.

6 **Please turn over:** When more risks follow on the next page, this is marked by a ***grey "turn over" icon***

7 **Please turn over:** When more mitigations follow on the next page, this is marked by a t***eal "turn over" icon***

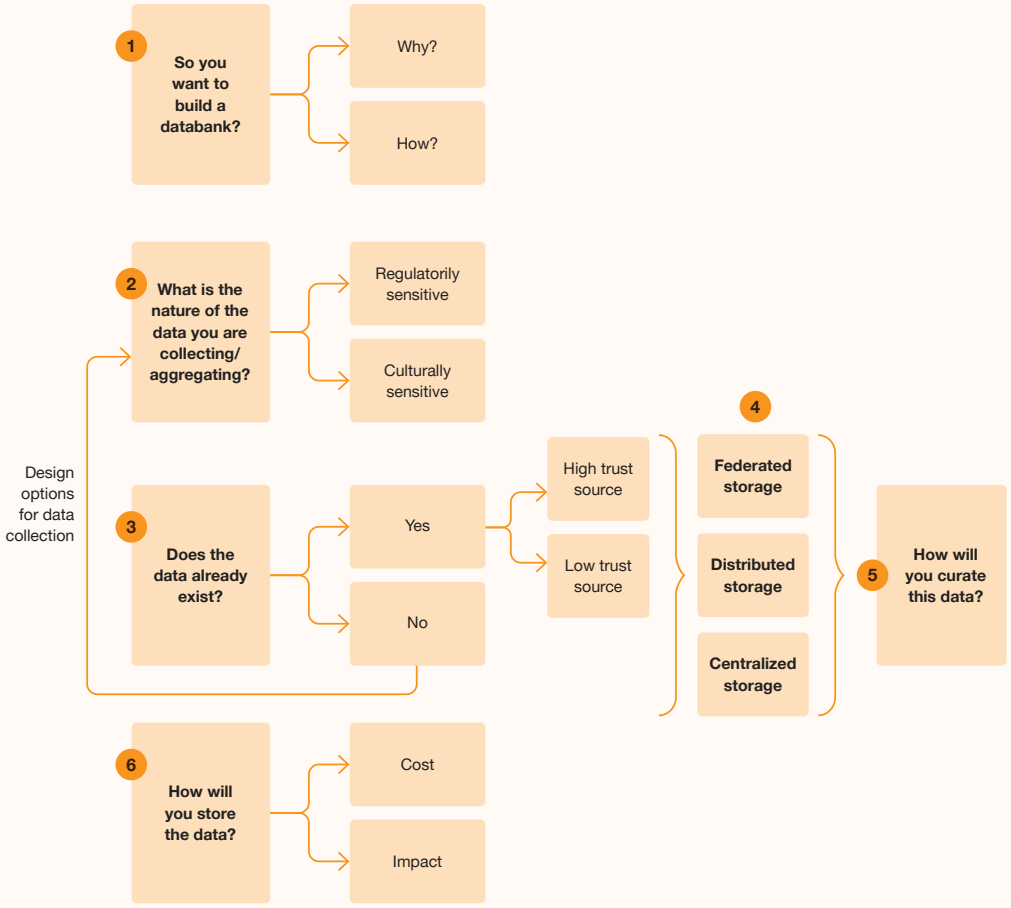
Glossary

Data grab	Large-scale gathering of information without meaningful consent e.g., about users of a website
Data type	Categorisation of data based on nature or source of data
Databank	A large collection of digital health information and biosamples drawn from many datasets
Databank builder	The primary entity that is developing and governing the databank
Dataset	Collection of data from a single source or intended for a single project
Derivable value	Generalised value that can be generated from research activities
Distributed value	Value realised by communities from research activities
LMIC	Low and middle income countries
HIC	High income countries
Participant or data subject	Individual who has contributed their data to the data bank
Participatory research at scale	Technical, design, and governance tools, features, and approaches that enable participatory research in big health data contexts
Platform hygiene	Technical, design, and governance tools, features, and approaches that encourage transparent, reproducible, inclusive science

Click to explore Phase 1 further

PHASE 1

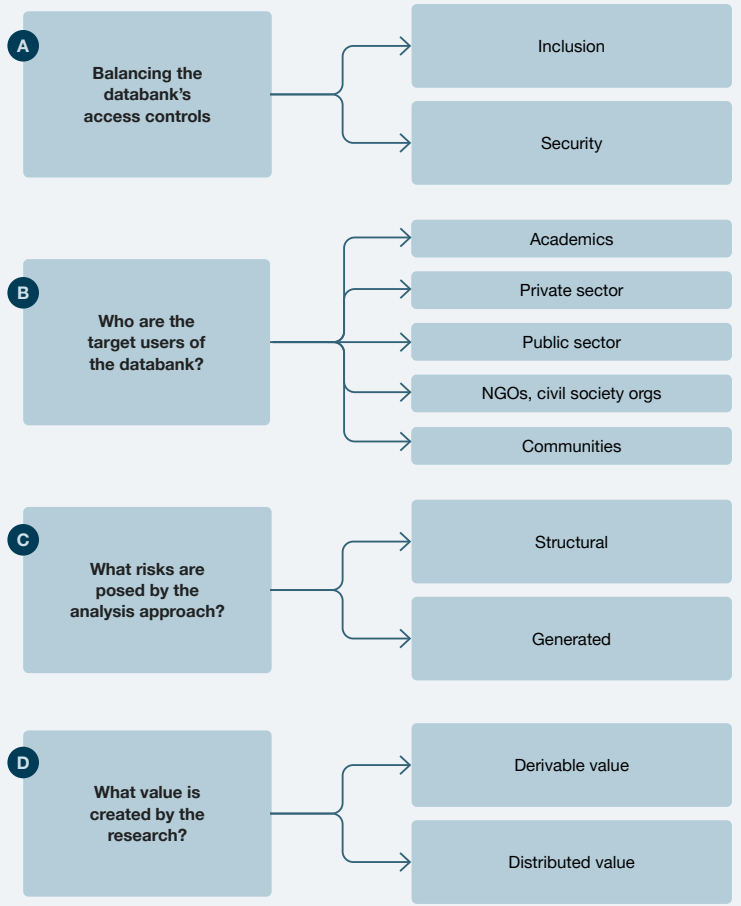
Developing a databank



Click to explore Phase 2 further

PHASE 2

Using the databank

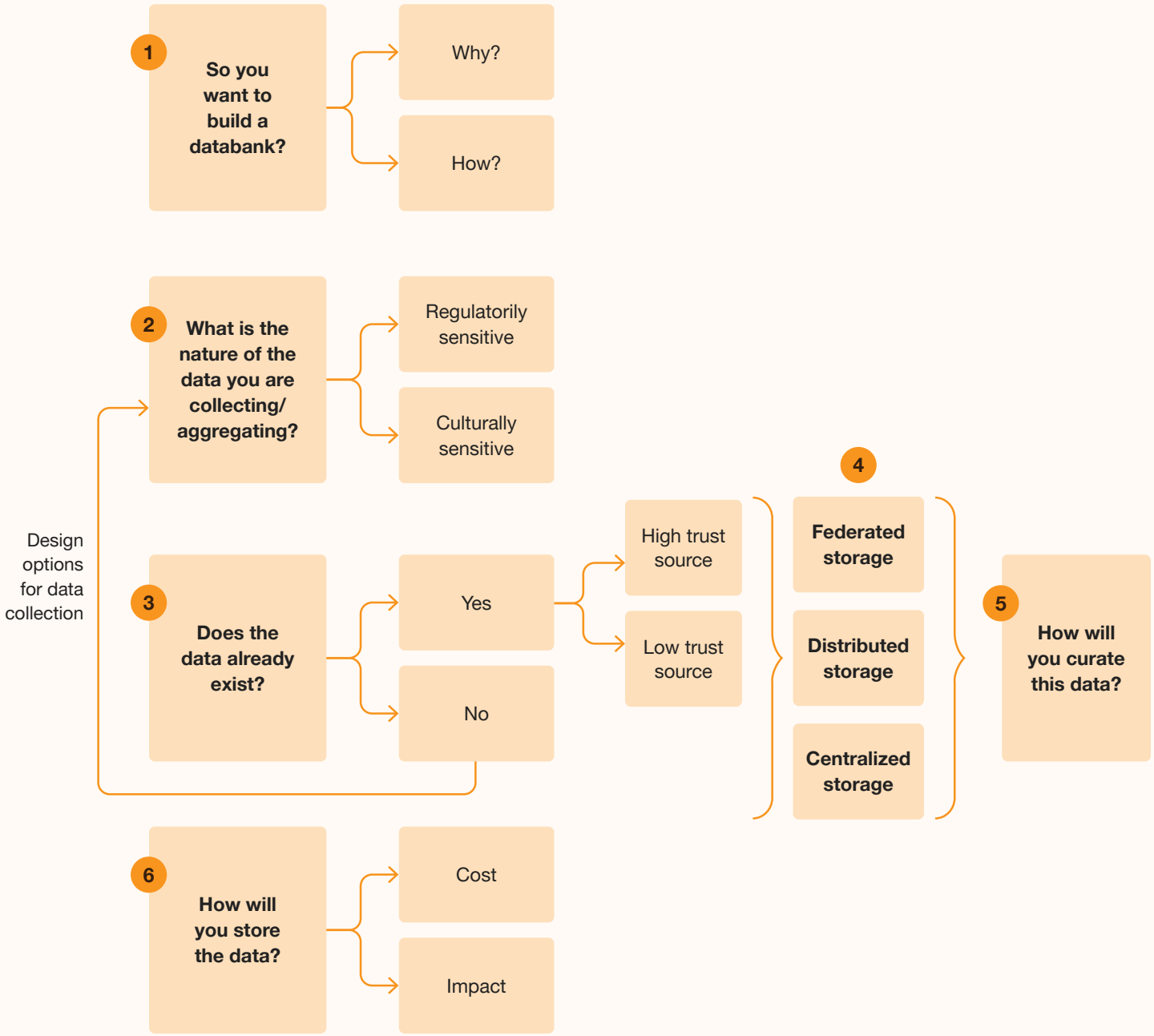


CONTENTS

Phase 1

Developing a databank

Click on the sections to navigate the document



CONTENTS

PHASE 1

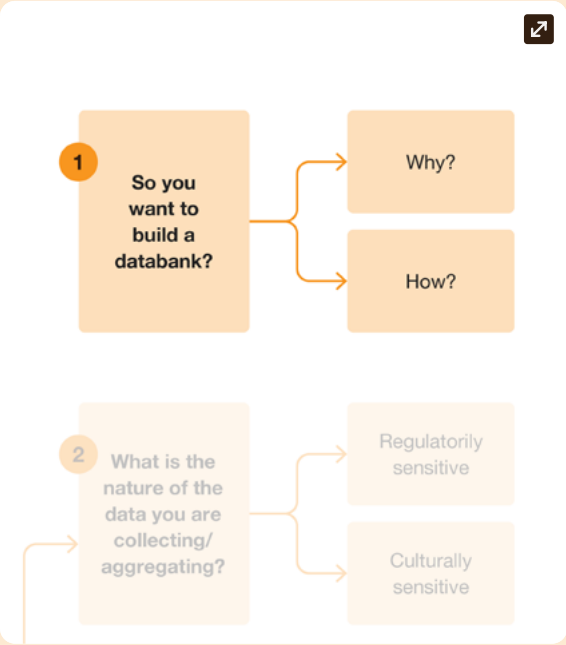
PHASE 2

DESIGN

1 So you want to build a databank?

PHASE 1

Developing a databank



RISKS	MITIGATIONS
Possibility of uninformed/poorly informed requisitioning of large amounts of data (“data grab”).	Decide, in consultation with communities/people with lived experience, if a given data acquisition is worth the possible solutions that it will help produce: what potential does the data have to create value for affected communities?
	Landscape awareness: Are other people collecting/using these data already? If yes, is it possible to do this research with minimal data collection and more reuse?
Environmental impact of big data collection, use, and storage.	For a given data cache, conduct an environmental impact assessment on that data’s collection, use, and storage.
	Conduct a landscape comparative to glean if similar analyses have been conducted with lower computational resources and environmental impact. Ensure use of the most up-to-date mechanisms for efficient computation.
	Limit large-scale computation to time-based cycles (e.g., particular quarters within the year), reducing overall annual usage.

DESIGN

So you want to build a databank?

PHASE 1

Developing a databank

Privacy protection comes at a cost to scientific solving and/or return of value.

See *"Data collection: regulatory and cultural considerations"* for mitigations associated with privacy. **Refer to page 10.**

Scientific hype cycle can add urgency/ pressure that isn't actually there and push researchers.

Implicit prioritisation of HIC research interests over LMIC research interests.

Articulate ethical principles to guide all databank/ research enabling activities in collaboration with research and lived experience advisors.

Create processes and procedures to ensure decision making adheres to guiding ethical principles.

Conduct regular audits to assess the fit between the databanks' research activity decision making and guiding ethical principles.

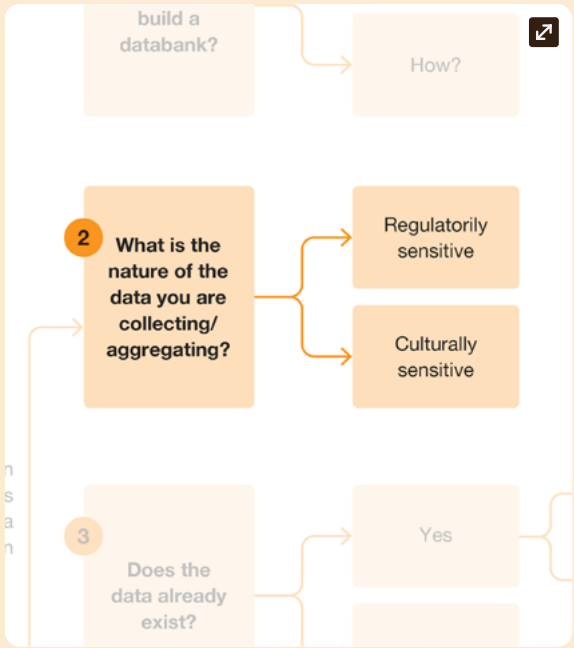
Support community/people with lived experience auditing the fit between the databank's guiding ethical principles/research activities and their own research interests.

DATA COLLECTION

2 What is the nature of the data you are collecting/ aggregating?

PHASE 1

Developing a databank



CONTENTS PHASE 1 PHASE 2

RISKS

MITIGATIONS

REGULATORILY SENSITIVE

Regulatory regimes may reinforce *performative* approaches to anonymization, ascribing greater value to anonymization and other privacy preserving approaches than is actually realised in practice.

Ensure any data/information about the databank that is publicly available is aggregated and de-identified.

Share information about data security processes and safeguard systems with participants and researchers: clear and complete disclosure of privacy protection, data security processes, and safeguard systems within the databank, ensuring transparency regarding the efficacy of the approach(es) to anonymization employed and other privacy protecting techniques.

- For existing datasets, build meaningful dialogue (scope could span from advisory board to public engagement) about the balance between the goals of data collection and use and attendant risks to privacy. Communities/people with lived experience must be involved in conversations weighing the risks and benefits of privacy protection and scientific solving. Consider structured, longitudinal methods e.g., Community Engagement Studios.
- For new data collection, in addition to community conversation regarding the balance of data collection and use/privacy, ensure informed consent is comprehensible and clear by conducting at least one round of user testing with prospective participants.

Platform hygiene Participatory research at scale

DATA COLLECTION

What is the nature of the data you are collecting/aggregating?

PHASE 1

Developing a databank

Engage participants in actively evaluating if their expectations for privacy protection for various data types or analyses matches the realised privacy protections for those data types/analyses (see for example Bridging the Gap PR specification 1 “Co-Creating and Implementing Community Safeguards, Security Tracking, and Data Security Explainer” p.29).

Engage participants and researchers in assessing the impact of local and contextual regulatory norms and regimes’ privacy requirements on scientific solving/community benefit.

Data *type* and *subject* definitions (and related rights/protections) may vary widely by jurisdiction, particularly in regard to protected groups and the types of data that can be collected from them, including “minors”.

Provide clear, easily understood definitions of key terms in a place readily accessible to communities, people with lived experience, participants, and researchers. Consider tying articles/resources to each definition in the glossary (see Bridging the Gap RH specification 4 “Definitions, Support Resources, and Research Stages” p. 25).

Document differences in protections through application of standard informed consent metadata (e.g., GA4GH’s DUO standard).

Establish within-databank norms to guide data acquisition/collection, access, and storage.

DATA COLLECTION

What is the nature of the data you are collecting/aggregating?

PHASE 1

Developing a databank

Co-create with communities/people with lived experience and research advisors a jurisdiction-agnostic code of protection and care for data subjects (i.e., participants) that applies to populations with low regulatory shielding (e.g., countries without data protection laws) as well those with high regulatory shielding (e.g., GDPR subjects).

Regulatory norms may impose localisation rules for different types of data which may affect centralisation, harmonisation, storage.

Develop and support infrastructure that enables a decentralised (but centrally controlled) data storage/use model and/or a fully federated storage/use model to allow for the greatest flexibility in data aggregation. NB: these approaches have important implications for access control, oversight, and will impact use (see the "Functional Requirements" section of the ISDA Rulebook).

Support and fund community owned and governed data storage infrastructures that can be part of the federated networks of data.

Some data types may have shorter shelf lives because of jurisdiction-specific regulation.

Build researcher and community-facing dashboards to chart the period of availability for certain data.

DATA COLLECTION

What is the nature of the data you are collecting/aggregating?

PHASE 1

Developing a databank

In consultation with communities, support periods of intensive engagement/outreach with researchers to focus research efforts on data with shorter “shelf life” corresponding with data collection/release cycles, e.g., through DREAM Challenges.

Data subjects may have limited or variable autonomy/decision making capacity to consent at point of data collection and/or variable autonomy/ decision making capacity over time and there are no standards for addressing variable autonomy over time in big data/secondary use research.

Explicitly assess and document participant autonomy/decision making capacity at the time of data collection; consider including advanced consent documentation (i.e., what to do should a participant’s capacity diminish or vary) or similar approaches.

Co-create with participants a databank-wide approach to assessing autonomy over time that allows participants to opt-in to a co-created, individually-specified “decreased autonomy workflow” for their data and build solutions such as temporary delegation of consent to care givers/guardians.

DATA COLLECTION

What is the nature of the data you are collecting/aggregating?

PHASE 1

Developing a databank

National security exceptions to data requests.

Data used to criminalise or marginalise people in state jurisdiction.

Use of data for surveillance of communities/ individuals/ targeted groups especially by government agencies.

Use of data to drive malicious or corrupt state interests.

Radical transparency regarding national security exceptions and other government use to data privacy and the implications to participants and communities/the public:

- For existing datasets, mitigations could span from engagement with communities/people with lived experience and scientific advisory boards to public engagement.
- For new data collection, additionally ensure exceptions are highlighted and potential implications described within all informed consent processes.

Periodically highlight this consideration/update participants on any known cases within this databank or others (e.g., this case from the US) and ensure adequate withdrawal/data destruction procedures.

Build internal/external legal capability to analyse global precedents on challenges to national security exceptions, and consider strategic litigations to challenge governments (if required) to create precedents.

DATA COLLECTION

What is the nature of the data you are collecting/aggregating?

PHASE 1

Developing a databank

RISKS	MITIGATIONS
Regulatory regimes may prioritise consideration of individual harms over potential harms to groups or exclude the consideration of group harms (as in the United States) entirely.	Document and disclose potential risks to groups (both claimed and imposed) posed by data aggregation activities in addition to doing so for individual participants.
In data-rich research, implicated "communities" includes not only self-identified or "claimed" communities but also imposed, algorithmically defined groups.	Track of emerging regulatory language on community data rights, assess impact on databank governance.

CULTURALLY SENSITIVE

Even in the absence of regulatory requirements, there still may need to be limits on the way data is collected, aggregated, and/or used for cultural reasons.	Engage local community experts to understand what cultural and contextual sensitivities may pertain to the data and to identify boundaries with regards to data use.
	Build community based infrastructure (technology, cultural experts) that supports iterative review/flagging for cultural sensitivities of data held by the databank/ planned for aggregation.

DATA COLLECTION

What is the nature of the data you are collecting/aggregating?

PHASE 1

Developing a databank

Participants' desire for adherence to cultural norms regarding biological sample (or other data) collection, storage, and destruction may impact the equity and inclusivity of the dataset over time.

Conduct ongoing monitoring and disclosure regarding the equity and inclusivity of the databank over time.

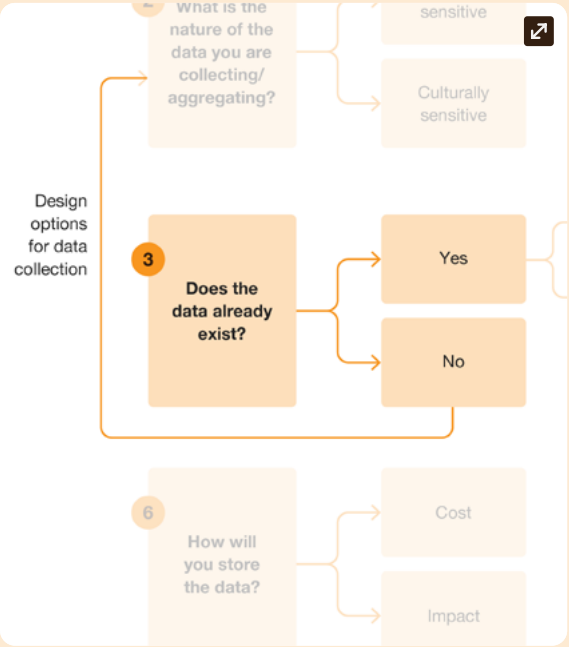
As needed, support remediation (e.g., through new sample collection), to ensure equity and inclusivity of the databank over time.

DATA COLLECTION

3 Does the data already exist?

PHASE 1

Developing a databank



RISKS	MITIGATIONS
YES, THE DATA ALREADY EXISTS	
Data is available only from a low trust source, e.g., data from corrupt state actors, misaligned private sector actors.	Establish databank goals for inclusivity of data with a focus on promoting equity.
Drive for creating a "representative" dataset could lead to pressure to aggregate data from low trust sources.	Create databank standards for "reputable data sources" in collaboration with research advisors and community/people with lived experience.
Acquiring data from a low trust source may perversely incentivise further extractive data processes.	Ensure community-facing transparency regarding decision-making to include/exclude a given dataset.
	Use contracting with data suppliers/data vendors as a lever for promoting equity and inclusion of datasets.
	Support a standing team of researchers and people with lived experience to weigh and advise on the benefits/downstream impact of data acquisition both for scientific advancement and impacted communities over time (e.g., via Community Engagement Studios).

DATA COLLECTION

Does the data already exist?

PHASE 1

Developing a databank

Misalignment of consent conditions when data is aggregated.

The conditions of consent under which the data were collected are unclear or suspect (e.g., click to agree, ToS, consent to data/sample collection as a condition of receiving healthcare).

Application of standard informed consent metadata (e.g., GA4GH's DUO standard) to allow for cross comparison of data caches' consent terms during aggregation.

Ensure all datasets considered for ingestion were legally collected (regulations vary by jurisdiction).

Support a standing team of researchers and people with lived experience to iteratively weigh and advise if a given dataset is sufficiently valuable to allow for its ingestion (e.g., via Community Engagement Studios).

Data exists but is analogue.

Publicly acknowledge data gaps resulting from existing but offline datasets.

Through engagement with governments and communities/people with lived experience identify analogue datasets (regional governments, nonprofits) relevant for the databank and fund their digitalisation.

DATA COLLECTION

Does the data already exist?

PHASE 1

Developing a databank

RISKS	MITIGATIONS

NO, THE DATA DOES NOT ALREADY EXIST, PROMPTING DATA COLLECTION

Risk of inadequate data collection (inequitable, exclusive).

Ongoing gap analysis of equity and inclusivity of dataset and evaluation of dataset against benchmarks of performance with research experts and communities/ people with lived experience.

Build and support infrastructure that brings community members into conversation with researchers to identify gaps in equity/inclusivity of dataset.

Data type is not available equally from all contexts because of different access to technology (e.g., wearable data from youth in LICs).

Data collection cost/sustainability may vary widely across contexts.

Data may be pragmatically challenging to obtain and these challenges may not be equal across all jurisdictions. For example, if target data is in medical records, its extraction may be near impossible if those records are paper, stored in a conflict zone, etc.

Document and openly disclose differences in availability of various data types.

Establish strong relationships with local research partners to help flag and mitigate the risks associated with unavailability of data and identify mitigation approaches. For example, when commissioning data collection consider providing the data collection device (e.g. FitBit, smartphone, etc.) rather than only recruiting participants who already have it.

DATA COLLECTION

Does the data already exist?

PHASE 1

Developing a databank

Ensure the technology (i.e., apps and other technology) used in data collection are inclusive of a global audience and built for interoperability.

- Adapted to different types of devices (smartphones, tablets, computers, etc.) as well as older models of these devices.
  - Text and images should be supported on a variety of devices especially small screens, with ability to zoom in and change text size.
  - Lean toward making materials image-rich (rather than text-rich) given blocks of text can be overwhelming on small screens.
- Adapted to different levels of reliable infrastructure (electricity and internet access).
  - Provide versions that work on low-speed connections.
  - Lean toward asynchronous (rather than synchronous) use options to allow for unpredictable internet access/electricity.
- Built for interoperability.
  - Collect new data using global standards for interoperability (e.g. HL7 CDA).

Ongoing assessment of the impact of differences in availability on research scope, applicability of insights, potential benefit to communities in collaboration with research experts and communities/people with lived experience.

DATA COLLECTION

Does the data already exist?

PHASE 1

Developing a databank

Lack of awareness and literacy among participants regarding impact of data collection (i.e., people don't know what they are giving away).

For new data collection, ensure informed consent is comprehensible, comprehensive, culturally tailored, and clear as a condition of funding.

Drive for creating a "representative" dataset could lead to coercion in collection e.g., through use of disproportionate incentives.

Document challenges faced in informed consent processes to report back to databank, and devise methods to overcome these with community/lived experiences advisors.

Lack of specific, unambiguous informed consent obtained from participants.

Ongoing reporting to advisors with lived experience on data collection protocols, with systems in place that empower these advisors to raise a flag if procedures are of concern.

With the guidance of local experts, advisors with lived experience, and research advisors balance the coercive potential of participation incentives/compensation against considerations of fairness, equity, and access (see MindKind Study Final Report p.201).

Fund local capacity building in advance of data collection; data collection teams should ensure strong partnership/co-leadership with local community representatives to assess needs for/ tailor capacity building.

Require/fund adequate supports to allow for truly informed consent in all contexts and conditions e.g., in person consent, interactive consent.

DATA COLLECTION

Does the data already exist?

PHASE 1

Developing a databank

The research demands of a databank are not conducive to individual participant sign off on particular secondary uses (i.e., granular consent).

Acknowledge the limitations of consent (i.e., if/when/why reconsent is not feasible) in the context of secondary usage, as applicable.

See Bridging the Gap for approaches to community engagement at scale (especially PR specification 3's "Dedicated Area for Feedback" and "Public Draft of Analysis" p.32 and PR specification 6 "Request a Brainstorm" p.40).

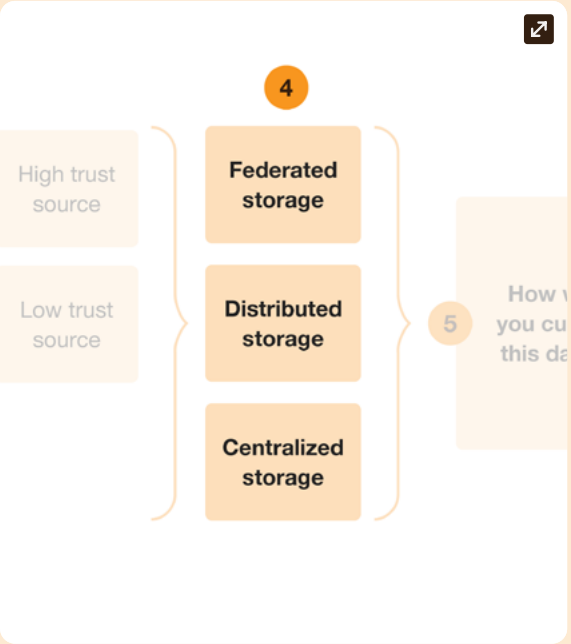
Implement infrastructure that requires researchers/access approvers to positively consider group interests/guard against the potential for group harm resulting from their work (e.g., Sage Bionetworks' Community Consent Toolbox).

DATA AGGREGATION

4 How will you aggregate this data?

PHASE 1

Developing a databank



RISKS	MITIGATIONS
As data are aggregated for secondary use, they are further removed from their community context.	See Bridging the Gap for ideas for approaches to community engagement at scale, especially "PR specification 3: Expert Advice, Dedicated Area for Feedback, and Extensions: Public Draft of Analysis, Field Notes" p.32 and "PR specification 6: Request a Brainstorm" p.40.
The demands of data scale and speed of databank building drives partnership with technology platform vendor(s) that may have low/lower public trust.	Through vetting and meticulous contracting with all databank enabling partners that incorporates oversight and guidance by advisors with lived experience and is consistent with the ethical principles of the databank.
Platform vendor as source of non-compliance with the spirit of data regulations (i.e., legal compliance in the absence of ethical, cultural, or moral compliance).	Through vetting, meticulous contracting, and value-aligned incentive structure (e.g., ensuring funding tied to outcomes in addition to just legal compliance) with all databank enabling partners that incorporates oversight and guidance by advisors with lived experience and is consistent with the ethical principles of the databank.

DATA AGGREGATION

How will you aggregate this data?

PHASE 1

Developing a databank

Difficult to institute community decision making because access already may be determined by data broker/set at collection.

Make visible pathways of access to communities to equip with information on who is viewing/ using data related to them (See Bridging the Gap tools such as RH specification 5 “Global Tracker & Progress Updates” p.26).

Consult with advisors with lived experience to determine if there are meaningful pathways for community engagement and, if not, if the data are still worth aggregating.

Develop and implement infrastructure that allows for researchers and participants to collaborate on research ideas as a mitigation approach for lack of community decision making power in access due to preset access. See Bridging the Gap tools such as PR specification 3’s “Dedicated Area for Feedback” p.32 and PR specification 6 “Request a Brainstorm” p.40.

If access controls are not universal for all data within the databank (e.g. due to federation or specific access requirements set at data collection) some data may become overused or hypervisibilised which can result in exploitation and/or inequity/exclusion.

Audit which datasets are most frequently used, review with advisors with lived experience and research advisors to identify concerns, implications for equitable solving.

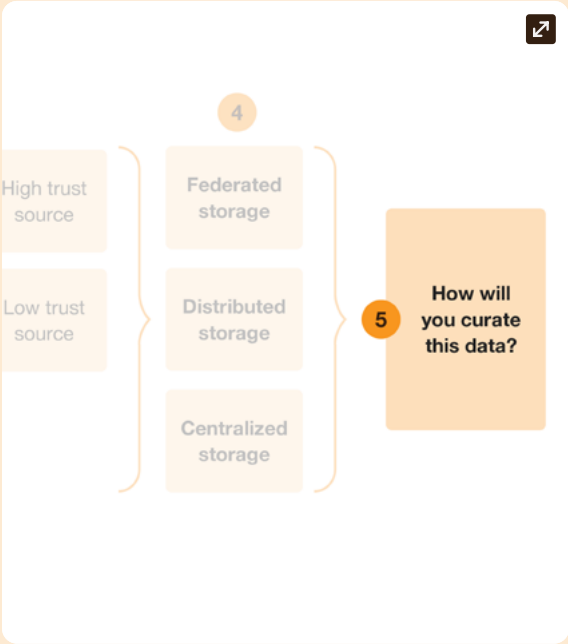
In collaboration with communities, advisors with lived experience, and scientific advisors, develop and implement mitigation approaches to address inequity caused by overused or hypervisibilised data.

CURATION

5 How will you curate this data?

PHASE 1

Developing a databank



RISKS	MITIGATIONS
When data is collected using standard measures across diverse cultures, equivalent data values may not have equivalent meaning (e.g., due to differences in diagnostic criteria, in diagnosis rates, in access) (for example of adaptation see Malawi Longitudinal Study of Families and Health).	Consult with local community experts to guide reconciliation of data labelling schemes.
	Tag datasets with contextual description of those data (e.g., Data Nutrition Project).
Data dictionary may not align with different community groups' self-definitions.	Develop standard metadata terms that can be tied to data elements to present contextual description in collaboration with lived experience advisors/community experts.
Choice of data harmonisation strategy for data from varied origins (geographic, commercial) implicitly prioritises some cultural norms over others.	Identify and apply metadata standards that support contextual description of individual data elements; if none exist, support their community-engaged codevelopment.
	Develop and support infrastructure that engages researchers and participants in determining how best to harmonise meanings across contexts (see for example Bridging the Gap PR specification 4's "Co-Creating Definitions" p.35).

CURATION

How will you curate this data?

PHASE 1

Developing a databank

Data generated for commercial purposes may be harmonised/ curated to different data standards than research data.

Audit of data standards; transparent selection in consultation with advisors with lived experience, including releasing a public impact statement.

There are no data harmonisation or curation standards that are community informed.

Ensuring adequate budget (fiscal, time, resource) to fulfil curation goals over time.

Identify pre-existing benchmarking standards that prioritise equity (e.g., Gender Shades).

Codevelop new data standards with communities/people with lived experience and research advisors (consider leveraging groups like GA4GH).

Develop and support infrastructure that engages researchers and participants in determining how best to harmonise meanings across contexts (see for example Bridging the Gap PR specification 4's "Co-Creating Definitions" p.35).

CURATION

How will you curate this data?

PHASE 1

Developing a databank

Curation of some forms of data (e.g., biological samples) may need to be localised due to regulation; cost of this curation may limit inclusion of data from lower resourced contexts.

Ensure sufficient support (financial, staffing, infrastructure) for curation in LMICs to support equity, inclusivity of databank.

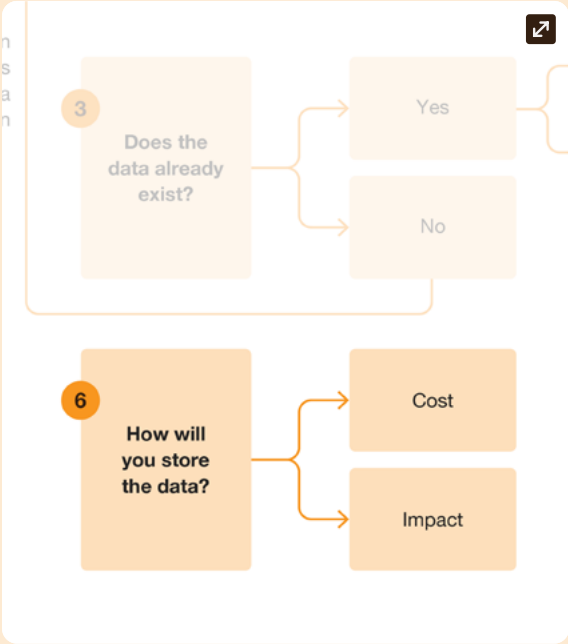
Consult with local communities/people with lived experiences to understand risks of misuse of localised data e.g. breaches of data centres, misuse by government and identify workarounds.

DATA STORAGE

6 How will you store the data?

PHASE 1

Developing a databank



RISKS	MITIGATIONS
Databank builder bearing data storage fees may unequally incentivize some data holders to allow for centralization/ cede local control.	Minimise financial incentives to relinquish data by equally funding both centralised and local storage schemes.
	Engage local data subject (in addition to local data holders) in decision making to centralise/ retain local control.
Regulatory or cultural restrictions on storage may reduce inclusion of some groups (e.g., indigenous communities) due to removal of samples.	Conduct complete assessment of regulatory and cultural requirements for data retention/ recontact prior to taking on storage, including impact assessment of resulting data/sample attrition in cases of lost contact.
If regulations or cultural norms require periodic re-consent or recontact, not all data/sample contributors will be equally reachable, especially people who lack stable housing, people living in areas with unstable infrastructure.	Develop and publicly share plans for data and sample retention that are appropriately tailored based on regulatory and cultural requirements.
	Adequately plan for and support recontact efforts.
	Conduct ongoing monitoring and disclosure regarding the equity and inclusivity of the databank as samples are removed.

DATA STORAGE

How will you store the data?

PHASE 1

Developing a databank

As needed, support remediation (e.g., through new sample collection), to ensure equity and inclusivity of the databank over time.

Environmental impact of data storage.

See *"Design"* for mitigations associated with environmental impact. **Refer to page 8.**

The costs (maintenance, compliance) of storage of primary data/samples may impact sustainability of databank as a whole and must be balanced against the potential benefit of storing primary data/samples for reinterpretation over time.

Institute joint decision making regarding data/sample storage with research advisors, local experts, and people with lived experience; share decision making approach and results openly.

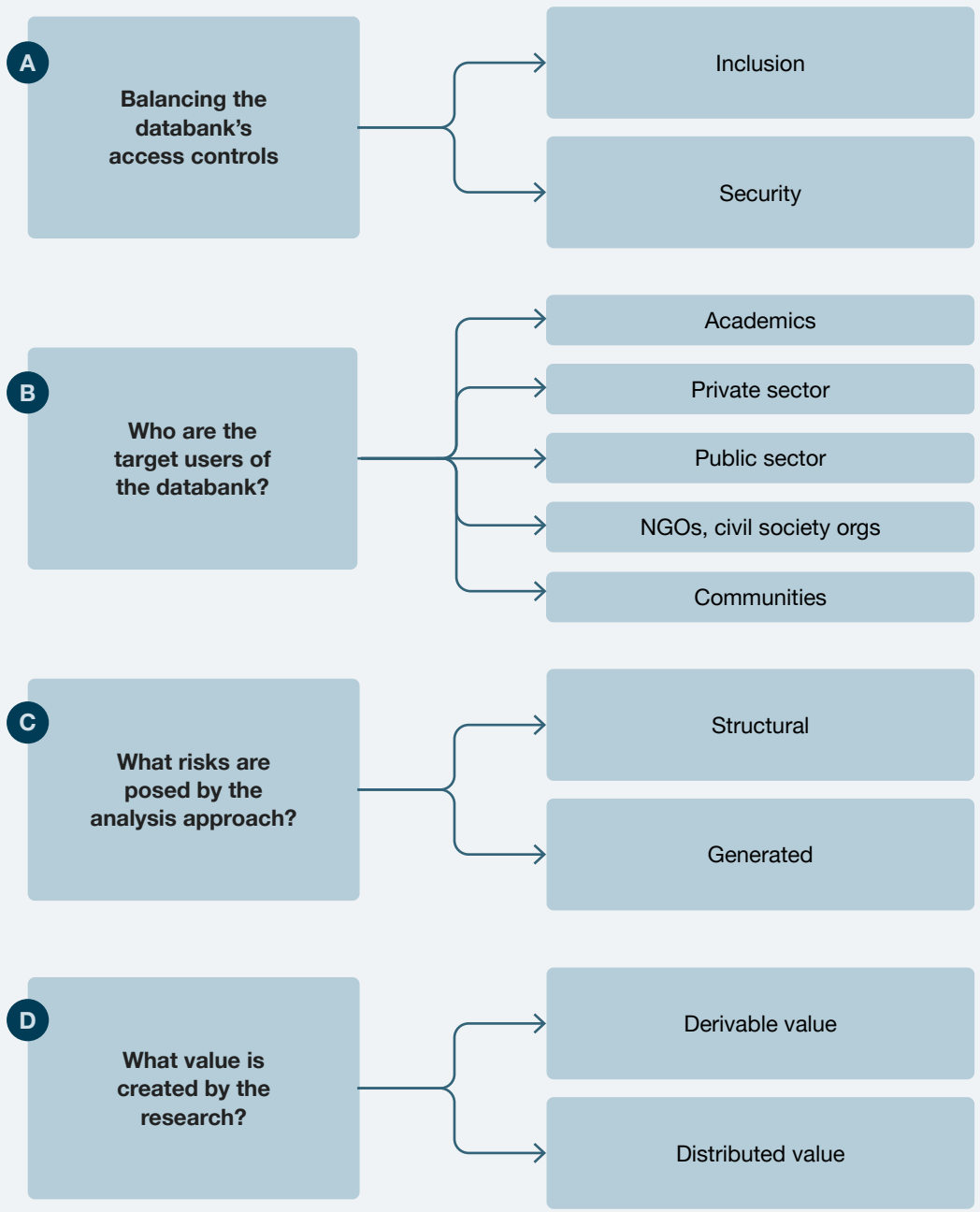
Institute and support standard terms of data retention across the biobank allowing for sunsetting of data for compliance/environmental/fiscal reasons.

If data are not centrally held, long term storage costs may be borne inequitably.

Phase 2

Using the databank

Click on the sections to navigate the document



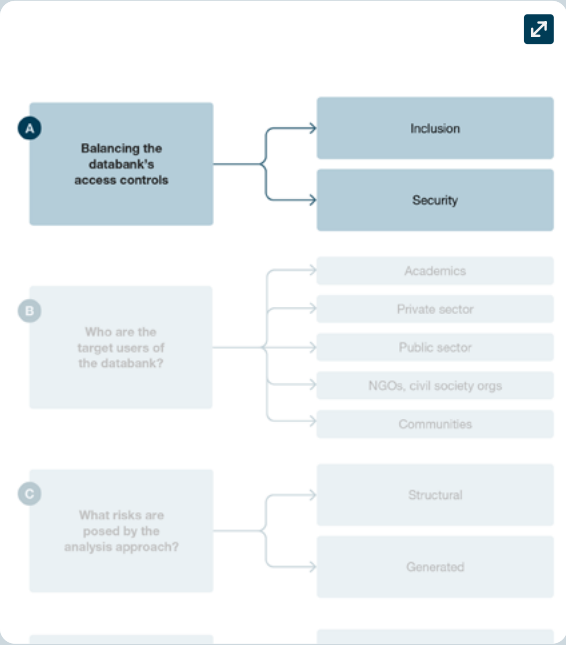
CONTENTS
PHASE 1
PHASE 2

ACCESS CONTROLS

A Balancing the databank's access controls

PHASE 2

Using the databank



RISKS

Access controls may block/prioritise some researchers because of bandwidth requirements, credentialing (e.g., US eRA Commons ID) requirements, and/or external ethics review requirements.

Having low access controls for certain requesters (private, state actors) may result in misuse of data.

Risks assessment models for providing access to requesters may be based on biases.

MITIGATIONS

Implement a nuanced system of credentialing that is inclusive of the broadest spectrum of researchers including citizen scientists (e.g., Sage Bionetworks' qualified researcher program) balancing against enabling access of "false flag" users.

Consult with lived experience advisors, local researchers, and others with contextual knowledge of barriers to data access in designing the databank user credentialing system.

Establish a databank governance body dedicated to oversight of researcher access and data use, including responsibility for periodic audit of barriers to access.

Establish a free-to-use databank ethics committee to address any ethics review requirements that the databank's data access committee wishes to impose.

Build and support infrastructure for community "sense checking" of research approaches/ outcomes (e.g., Bridging the Gap, PR Specification 3 "Public Draft of Analysis" "Field Notes" p.32 and PR specification 6 "Request a Brainstorm" p.40).

DATA STORAGE

Balancing the databank's access controls

PHASE 2

Using the databank

Lack of recourse in case of misuse of data.

Require data users to get institutional backing prior to data use (e.g., institutional data use agreement).

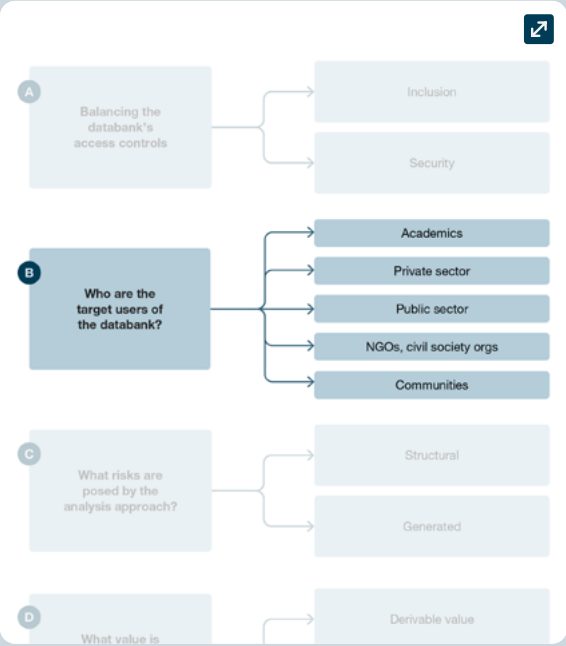
Develop a system of direct researcher licencing/bonding, modelled after existing systems for tradespeople.

DATABANK USERS

B Who are the target users of the databank?

PHASE 2

Using the databank



RISKS

The databank is underused by target users (i.e., databank not fulfilling its ethical obligation to participants).

Databank users drawn exclusively from the pool of people who are already known.

MITIGATIONS

Identify and document who target users of the databank are in collaboration with research and lived experience advisors.

Plan for ongoing databank support (budget, expertise, staffing) of key tenets of the *science of team science*, such as adequate researcher portal infrastructure and resources that support open communication and collaboration.

Host solving events, e.g., DREAM Challenges, to raise awareness of and bring solvers into the databank.

Expand the scope of the science of team science to include people with lived experience and other "non-traditional" solvers (for example as described in Bridging the Gap PR specification 1 "Co-Creating and Implementing Community Safeguards" p.29).

DATABANK USERS

Who are the target users of the databank?

PHASE 2

Using the databank

Inequitable distribution resources and capacity leads to inequity in data use.

Ensure transparency in databank use by monitoring and publicly sharing the number and variety (role, context) of users of the databank/ comparing with databank targets.

Databank infrastructure is exclusionary due to resource/ capacity requirements for its use.

Ensure databank infrastructure works equally well in high and low resource settings and/or drive equity through direct support of lower resourced researchers' databank use (e.g., support for computation resources, high speed internet costs, hardware).

Ensure the databank's infrastructure builds skills and knowledge (e.g., Bridging the Gap RH specification 3 "Education About Research, Ease of Navigation" p.24 and RH specification 4 "Definitions, Support Resources, and Research Stages" p.25).

Earmark funds for capacity building of LIMC and other disadvantaged context researchers.

Staff a helpline (e.g., chat, email, phone) to support new users/novice users of the databank.

Ensure equitable databank access controls *(see "access controls" for access control mitigations.* ***Refer to page 34****).*

DATABANK USERS

Who are the target users of the databank?

PHASE 2

Using the databank

Encourage or require researchers work in collaboration with participants/people with lived experience to identify and prioritise research foci (e.g., Bridging the Gap PR specification 3 “Field Notes” p35 and PR specification 6 “Request a Brainstorm” p.40).

Build infrastructure that supports non-traditional databank users’ (e.g., community members’) growth and recognition as researchers (e.g., Bridging the Gap PR specifications 1-6, especially PR specification 5 “My First Research Profile” p.38).

Risk of systemic inequities or exclusion of research subject groups being unwittingly furthered by well-meaning researchers.

Audit data usage to identify areas of underuse/overuse.

Incentivize use of data that has been underused through solving events, e.g., DREAM Challenges, and/or targeted support (budget, expertise, staffing).

Build and support infrastructure for community “sense checking” of research approaches/ outcomes (e.g., Bridging the Gap, PR Specification 3 “Public Draft of Analysis” “Field Notes” p.32 and PR specification 6 “Request a Brainstorm” p.40).

DATABANK USERS

Who are the target users of the databank?

PHASE 2

Using the databank

Lack of transparency regarding use of data.	Require researcher profiles to be publicly posted as a condition of databank use.
Potential for misinformed or biassed research design/agenda by data users.	Require research uses to be publicly posted as a condition of databank use with a flagging mechanism for community concerns (e.g., "request a review of this research project" feature in the All of Us Research Program's Data Hub).
Potential for use of well-meaning data analyses for targeted activities (selling, campaigning, etc) by malicious or misaligned actors.	Require plain language descriptions of research as a condition of data access (e.g., as text or video as in the All of Us Research Program).
	Require adherence in data use to applicable standards set by professional bodies (e.g., ASHG statement regarding concepts of "good genes").
There are malicious community actors with agendas.	Host ongoing dialogue regarding "acceptable" and "unacceptable" data uses with community/ people with lived experience recognizing that what is acceptable/unacceptable may evolve over time; implement this guidance as a component of databank access requirements.
	Build and support infrastructure for community "sense checking" of research approaches/ outcomes (e.g., Bridging the Gap, PR Specification 3, subsection "Public Draft of Analysis" p.32).

DATABANK USERS

Who are the target users of the databank?

PHASE 2

Using the databank

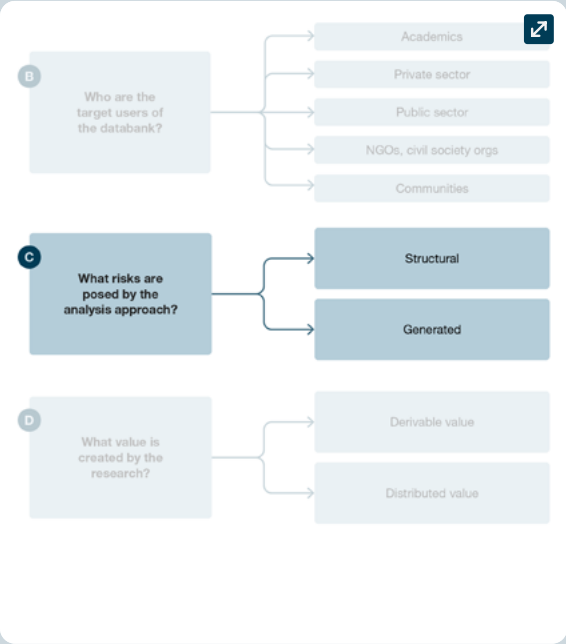
Mechanisms of research funding may skew which analyses are proposed and/or research design and/or methods of research.	Require disclosure of financial supporter(s) and terms of that support for each research project, link with research description and publicly post.
Funding behind a research request may affect research agenda.	Engage lived experience and research advisors to assess the overlap of research funders' priorities with community research priorities; use this information to supplement funding and/or for targeted communication with external funders.
Profit motives and shareholder interests may skew research design, or be covertly designed solely toward profitability.	

ANALYSIS

C What risks are posed by the analysis approach?

PHASE 2

Using the databank



RISKS	MITIGATIONS
Limitations on computational equity posed by cost of computation or local infrastructure (e.g., power grid stability, hardware) in low-resource settings.	Support through dedicated grants or other mechanisms computation costs/infrastructure to support researchers (e.g., supporting cost of computation, cost of hardware or internet access, or travel grants).
Bias in the dataset may result in bias generating/ perpetuating research and/or spurious findings.	See *"Curation"* for mitigations related to data curation. **Refer to page 25.**
Research may generate systemic harms and/or discriminate.	Build and support infrastructure for community "sense checking" of research approaches/ outcomes (e.g., Bridging the Gap, PR Specification 3, subsection "Public Draft of Analysis" p.32).
	Assess, adapt, and apply tenets of emerging systems of algorithmic reparation.

ANALYSIS

What risks are posed by the analysis approach?

PHASE 2

Using the databank

Researchers prioritising advantaged (HIC, white, global north) contexts over disadvantaged contexts.

See *"Data collection: existing and new"* for mitigations related to inequitable data collection. **Refer to page 17.**

Research does not recognize structural harms, systemic biases, historical contexts, and/or is unaligned with community desires.

Conduct ongoing monitoring and disclosure regarding the equity and inclusivity of the research done with the databank over time.

As needed, support remediation to ensure equitable creation of value of the databank over time through targeted databank governance (e.g., data use statement review), funding, and/or solving events, e.g., DREAM Challenges.

Implement a coproduction model for research agenda setting.

See Bridging the Gap for ideas for approaches to community engagement at scale, especially "PR specification 3: Expert Advice, Dedicated Area for Feedback, and Extensions: Public Draft of Analysis, Field Notes" p.32 and "PR specification 6: Request a Brainstorm" p.40.

ANALYSIS

What risks are posed by the analysis approach?

PHASE 2

Using the databank

Databank use for "diversity washing" of insights gleaned from homogenous populations.

Clearly define, and create processes for revisiting and renewing, the parameters of "diversity" and "representativeness" within the databank with a focus on promoting equity and inclusion (see, for example, GA4GH's "Diversity in Datasets Policy").

Require research uses to be publicly posted as a condition of databank use with a flagging mechanism for community concerns (e.g., "request a review of this research project" feature in the All of Us Research Program's Data Hub).

Require plain language descriptions of research as a condition of data access (e.g., as text or video as in the All of Us Research Program).

Develop an approach to disclosing potential data harms to encourage researcher's consideration of the potential (unintended) negative outcomes of their proposed research (e.g., Sage Bionetworks' Community Consent Toolbox).

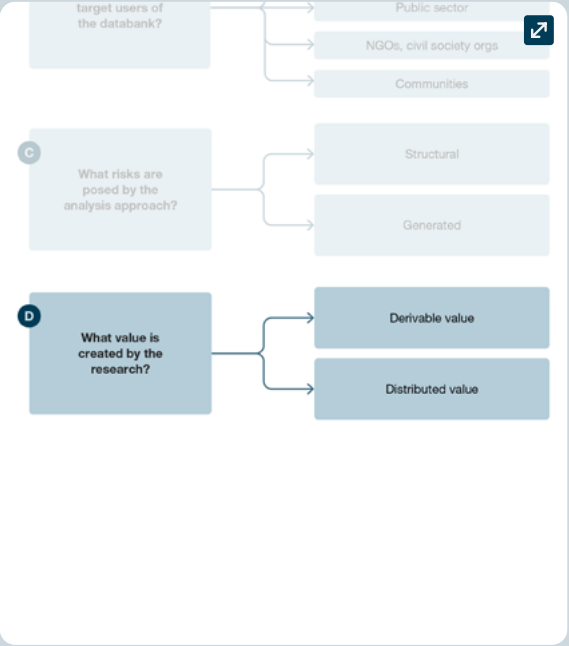
Build and support infrastructure for community "sense checking" of research approaches/ outcomes (e.g., Bridging the Gap, PR Specification 3 "Public Draft of Analysis" p.32).

RETURN OF VALUE

D What value is created by the research?

PHASE 2

Using the databank



RISKS	MITIGATIONS
Individual-level decisions on data use may, over time, disproportionately prioritise value toward certain stakeholders or disenfranchise others.	In collaboration with community experts/people with lived experience, define and apply a rigorous value framework to guide all databank governance decision making.
	Co-create varied and nuanced descriptions of 'value from research' in direct collaboration with communities/people with lived experience (effort could be supported by direct engagement using structured, longitudinal methods e.g., Community Engagement Studios or at a larger scale by reusing/repurposing technical infrastructure specified to enable Bridging the Gap PR specification 6 "Request a Brainstorm" p.40. Apply this 'value from research' framework to drive databank governance decisions.
Barriers to distribution of value from the databank, such as language, geography, and income, may undermine the value the research holds for communities.	Earmark sufficient resources (funding, expertise, staff) to support minimum standard processes return of value to communities (e.g., as described in Bridging the Gap RH specification 3 "Education About Research" p.24 and RH specification 4 "Definitions, Support Resources, and Research Stages" p.25, funding to cover open access publication fees for databank users).
Proper distribution of value among communities may be difficult due to the nature of value itself.	

RETURN OF VALUE

What value is created by the research?

PHASE 2

Using the databank

Engage with communities/people with lived experience to identify specific pathways and barriers to realising equitable value from key research initiatives (e.g., via Theory of Change as in this example).

Lack of accessible databank infrastructure may limit how much value communities can make of the databank as resource themselves.

See for *"Databank users"* for mitigations regarding databank accessibility.
Refer to page 33.

Acknowledgements

Authors

From Sage Bionetworks: Megan Doerr and Carly Marten

From Aapti Institute: Amrita Nanda, Rattanmeek Kaur, and Astha Kapoor

Design

From Sage Bionetworks: Stockard Simon

From Aapti Institute: Meher Rajpal and Antara Madavane

With thanks

We acknowledge with gratitude Wellcome staff and lived experience advisors, and external subject matter experts, including:

Wellcome staff:

Rebecca Asher
Hannah Atkinson
Matthew Brown
Kim Donoghue
Suzi Gage
Sarah Golding
Sophie Hawkesworth
Maisie Jenkins
Emily Jesper-Mir
Shomari Lewis-Wilson
Paul Meller
Dan Robotham
Winnie Wefelmeyer
Miranda Wolpert

Wellcome Lived Experience Advisers based in Australia, India, Indonesia, Japan, Kenya, Rwanda, South Africa, and the UK, including:

Dionisius Agnuza Jagadhita
Chantelle Booysen
Arkan Daffa Lazuardi
Grace Gatera
Meghna Khatwani
Jamie Morgan
Margaret Osolo Odhiambo
Tania Pandia
Dhriti Sarkar
Natasha Swingler
Veronica Wanyee

This work was funded by Wellcome.

About Wellcome

Wellcome supports science to solve the urgent health challenges facing everyone. We support discovery research into life, health and wellbeing, and we're taking on three worldwide health challenges: mental health, global heating and infectious diseases.

wellcome.org
contact@wellcome.org

About Aapti Institute

Aapti is a public research institute that works on the intersection of technology and society. We examine the ways in which people interact and negotiate with technology both offline and online.

aapti.in
contact@aapti.in

About Sage Bionetworks

Sage Bionetworks is a nonprofit health research organization that is speeding the translation of science into medicine. We believe that high-quality, well-annotated data acts as the foundation of modern biomedical innovation. We dream of a world where people work together across institutional boundaries to meaningfully address major medical research problems. We incubate new ways for diverse groups of people to practice research together.

sagebionetworks.org
megan.doerr@sagebionetworks.org

Produced by Sage Bionetworks and Aapti Institute for Wellcome Trust, 2023.