

Scalable Anonymization for Complex Clinical Trial Data

Abby McDonell, Sr. Program Manager





Problem Statement:

Clinical trial transparency is evolving beyond traditional data sharing in documents and tables to include complex formats such as imaging and genomic data. However, anonymizing these diverse data types in a unified, scalable way presents significant challenges; including the protection of privacy in linked data while maximizing data utility.

Multi-modal data sharing



Imaging

- Growing use in AI + secondary research
- De-identification is complex (metadata + “burned-in” text)
- Trend: standardized DICOM de-ID + validation workflows

Multi-modal data sharing



Imaging

- Growing use in AI + secondary research
- De-identification is complex (metadata + “burned-in” text)
- Trend: standardized DICOM de-ID + validation workflows



Wearables

- More remote, continuous endpoints
- Regulators now expect “fit-for-purpose” DHT selection + validation
- Trend: decentralized elements + hybrid trial designs

Multi-modal data sharing



Imaging

- Growing use in AI + secondary research
- De-identification is complex (metadata + “burned-in” text)
- Trend: standardized DICOM de-ID + validation workflows



Wearables

- More remote, continuous endpoints
- Regulators now expect “fit-for-purpose” DHT selection + validation
- Trend: decentralized elements + hybrid trial designs



Genomics

- High re-identification sensitivity + strict governance
- Shift toward controlled access + standardized terms/controls
- Trend: federated/“data visiting” models where feasible

Considerations for imaging-related data sharing

Increased demand

- Increased demand to share **MRI, X-ray, CT, ultrasound** alongside tabular datasets to give researchers a fuller, multimodal picture.
- Platforms like Vivli actively building pipelines to support **linked imaging + clinical data** sharing.

Considerations for imaging-related data sharing

Increased demand

- Increased demand to share **MRI, X-ray, CT, ultrasound** alongside tabular datasets to give researchers a fuller, multimodal picture.
- Platforms like Vivli actively building pipelines to support **linked imaging + clinical data** sharing.

Key challenges in image data anonymization

- Sensitive information exists in **two layers**:
 - DICOM metadata
 - **Burned-in pixel text** that may include PHI/identifiers
- Lack of global **standardization** across scanners, institutions, and formats makes risk patterns inconsistent.
- Anonymization must address both pixel data and metadata across **thousands of images**.

Considerations for imaging-related data sharing

Increased demand

- Increased demand to share **MRI, X-ray, CT, ultrasound** alongside tabular datasets to give researchers a fuller, multimodal picture.
- Platforms like Vivli actively building pipelines to support **linked imaging + clinical data** sharing.

Key challenges in image data anonymization

- Sensitive information exists in **two layers**:
 - DICOM metadata
 - **Burned-in pixel text** that may include PHI/identifiers
- Lack of global **standardization** across scanners, institutions, and formats makes risk patterns inconsistent.
- Anonymization must address both pixel data and metadata across **thousands of images**.

Emerging trends

- Momentum toward **standardized DICOM de-identification profiles**.
- Increased emphasis on **validation workflows**, not just de-ID execution (visual inspection + automated checks).
- Rapid growth in **AI-related use cases**, driving demand for large labeled imaging datasets—e.g., Gates Foundation ultrasound AI initiative for maternal health.

Ensuring readiness to meet increased demands for image data-sharing

1

Planning & resourcing

- Imaging adds significant prep time due to QC, burned-in text review, and metadata normalization.
- Early identification of imaging endpoints is critical, so anonymization isn't left to late-stage timelines.

Ensuring readiness to meet increased demands for image data-sharing

1

Planning & resourcing

- Imaging adds significant prep time due to QC, burned-in text review, and metadata normalization.
- Early identification of imaging endpoints is critical, so anonymization isn't left to late-stage timelines.

2

Technical & governance requirements

- Implement workflows for automated pixel-text detection + manual visual QC.
- Use consistent annotation mapping when linking images with tabular datasets.
- Establish imaging-specific SOPs covering ingestion, de-ID, QC, and error handling.

Ensuring readiness to meet increased demands for image data-sharing

1

Planning & resourcing

- Imaging adds significant prep time due to QC, burned-in text review, and metadata normalization.
- Early identification of imaging endpoints is critical, so anonymization isn't left to late-stage timelines.

2

Technical & governance requirements

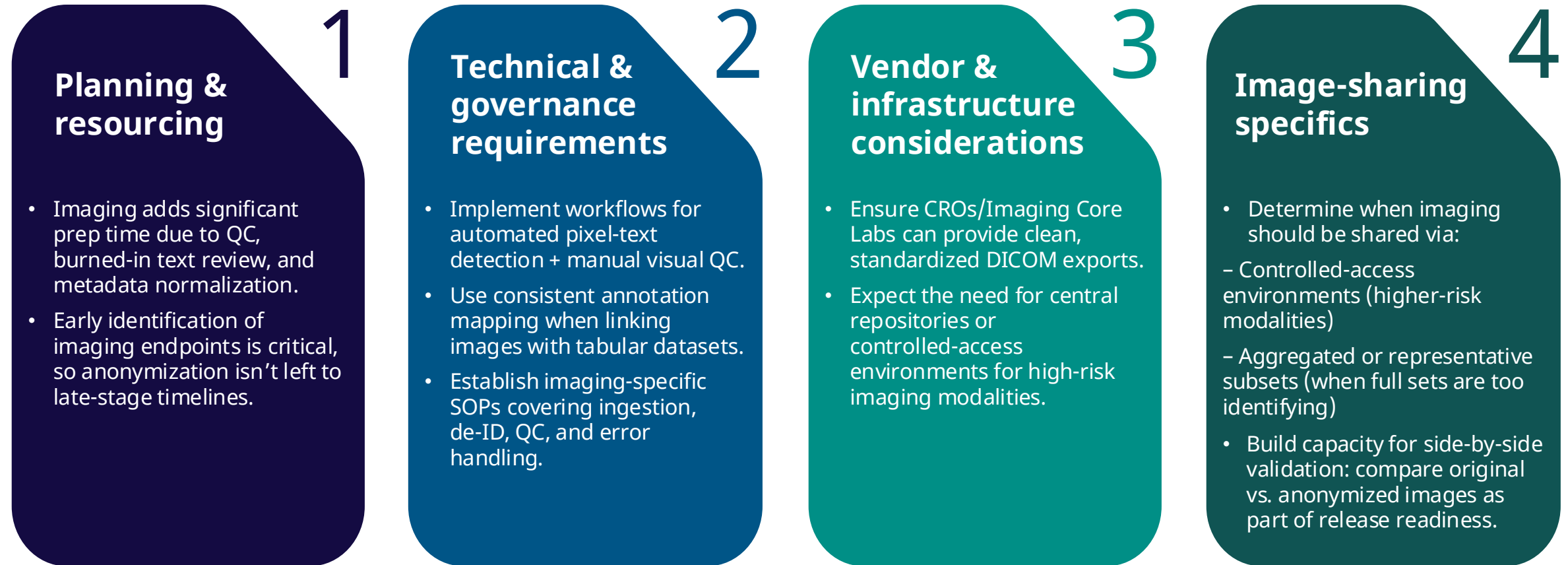
- Implement workflows for automated pixel-text detection + manual visual QC.
- Use consistent annotation mapping when linking images with tabular datasets.
- Establish imaging-specific SOPs covering ingestion, de-ID, QC, and error handling.

3

Vendor & infrastructure considerations

- Ensure CROs/Imaging Core Labs can provide clean, standardized DICOM exports.
- Expect the need for central repositories or controlled-access environments for high-risk imaging modalities.

Ensuring readiness to meet increased demands for image data-sharing



Considerations for genomic data handling



**Some data types
cannot be transformed
without destroying
utility, e.g.,**



Whole genome
sequencing



Fingerprint images



Retinal scans

Considerations for genomic data handling



**Some data types
cannot be transformed
without destroying
utility, e.g.,**



Whole genome
sequencing



Fingerprint images



Retinal scans



**Genomic data must be
combined with other
modalities for re-
identification**

e.g., referenceable,
identified data

Considerations for genomic data handling



Some data types cannot be transformed without destroying utility, e.g.,



Whole genome sequencing



Fingerprint images

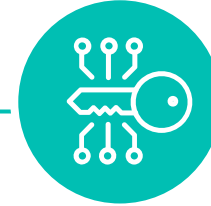


Retinal scans



Genomic data must be combined with other modalities for re-identification

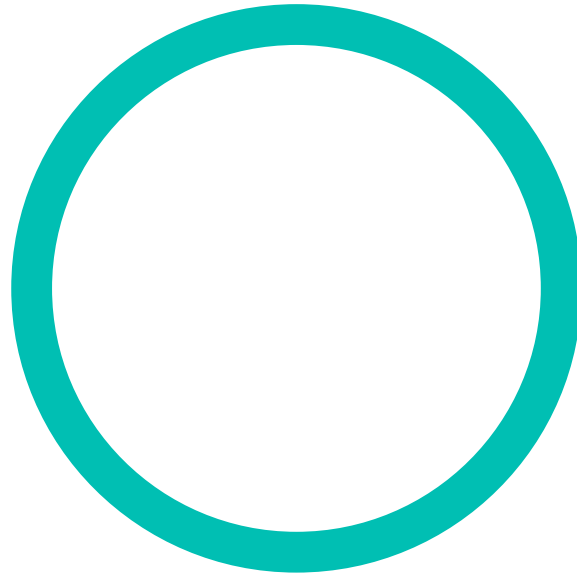
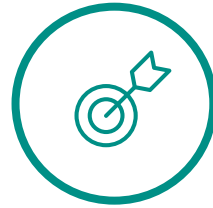
e.g., referenceable, identified data



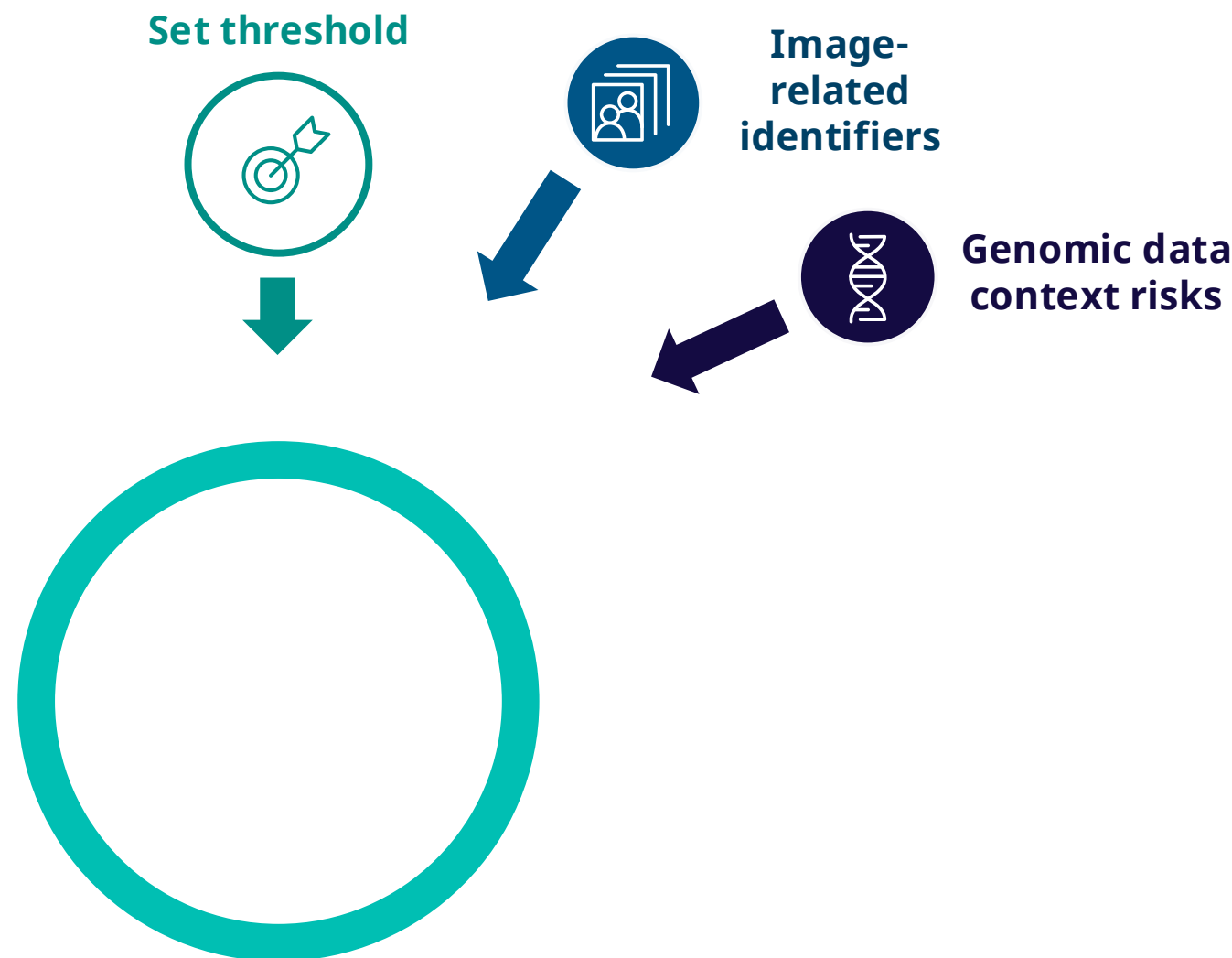
Risk mitigation must focus on **best-in-class context controls**

Statistical Anonymization Cycle

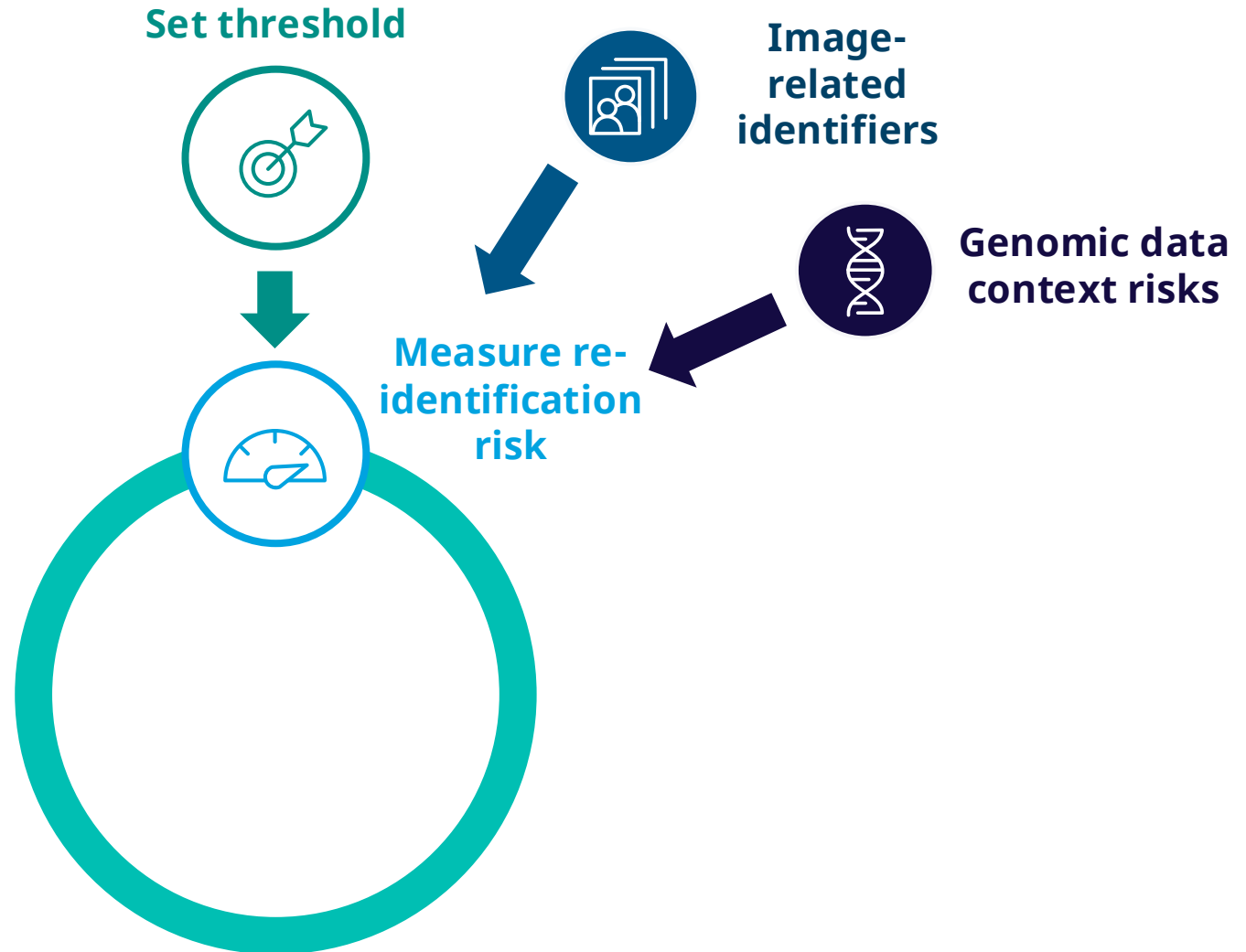
Set threshold



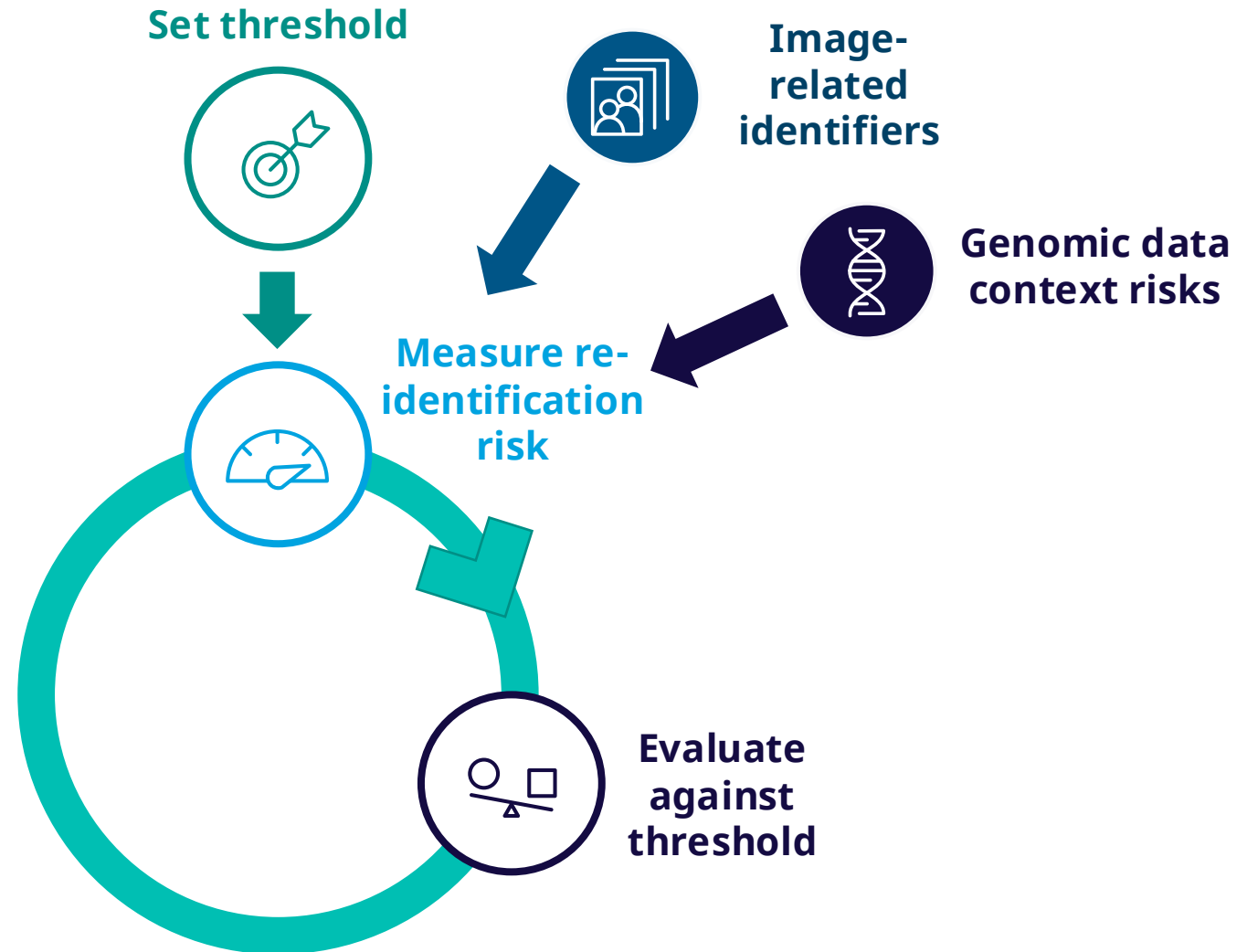
Statistical Anonymization Cycle



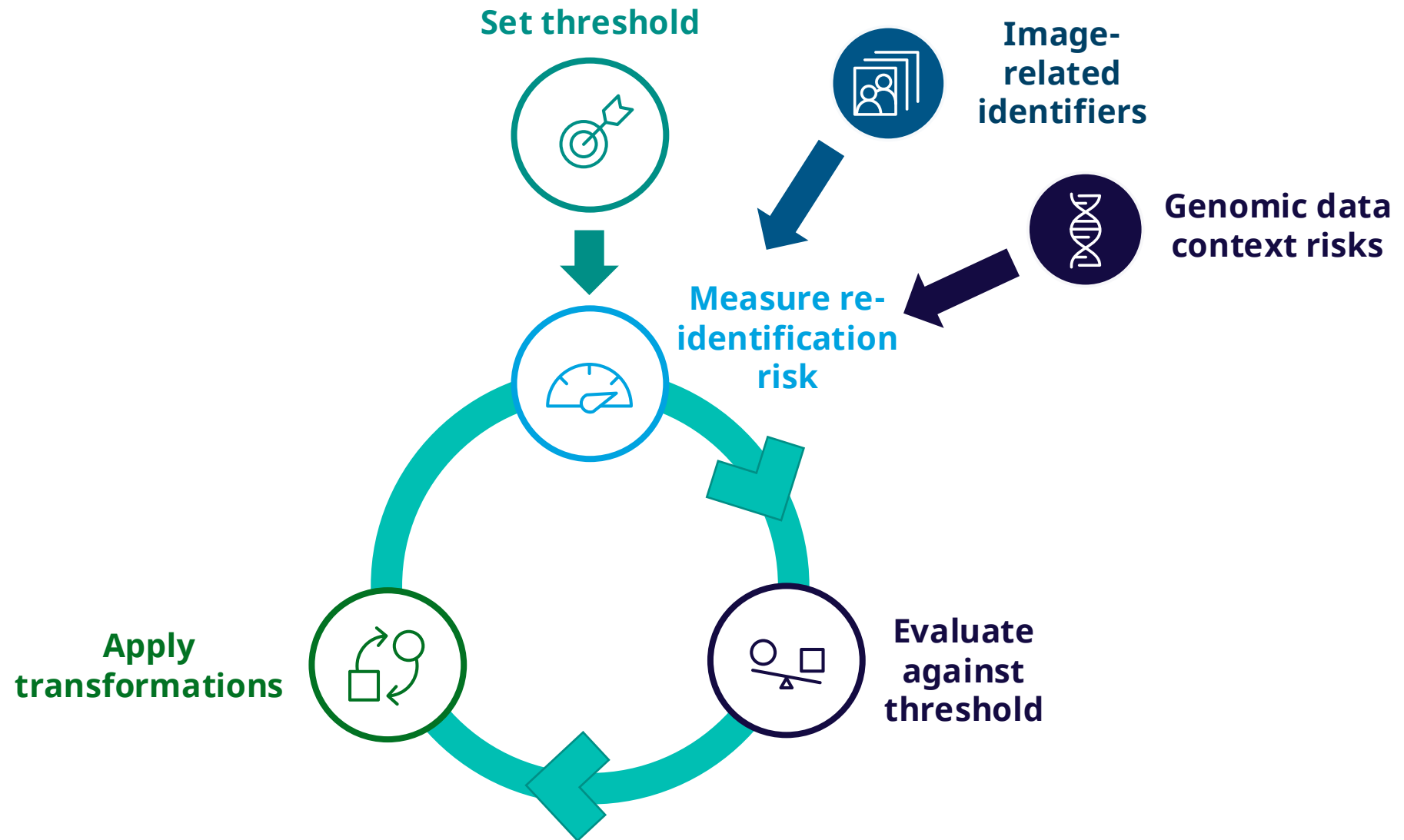
Statistical Anonymization Cycle



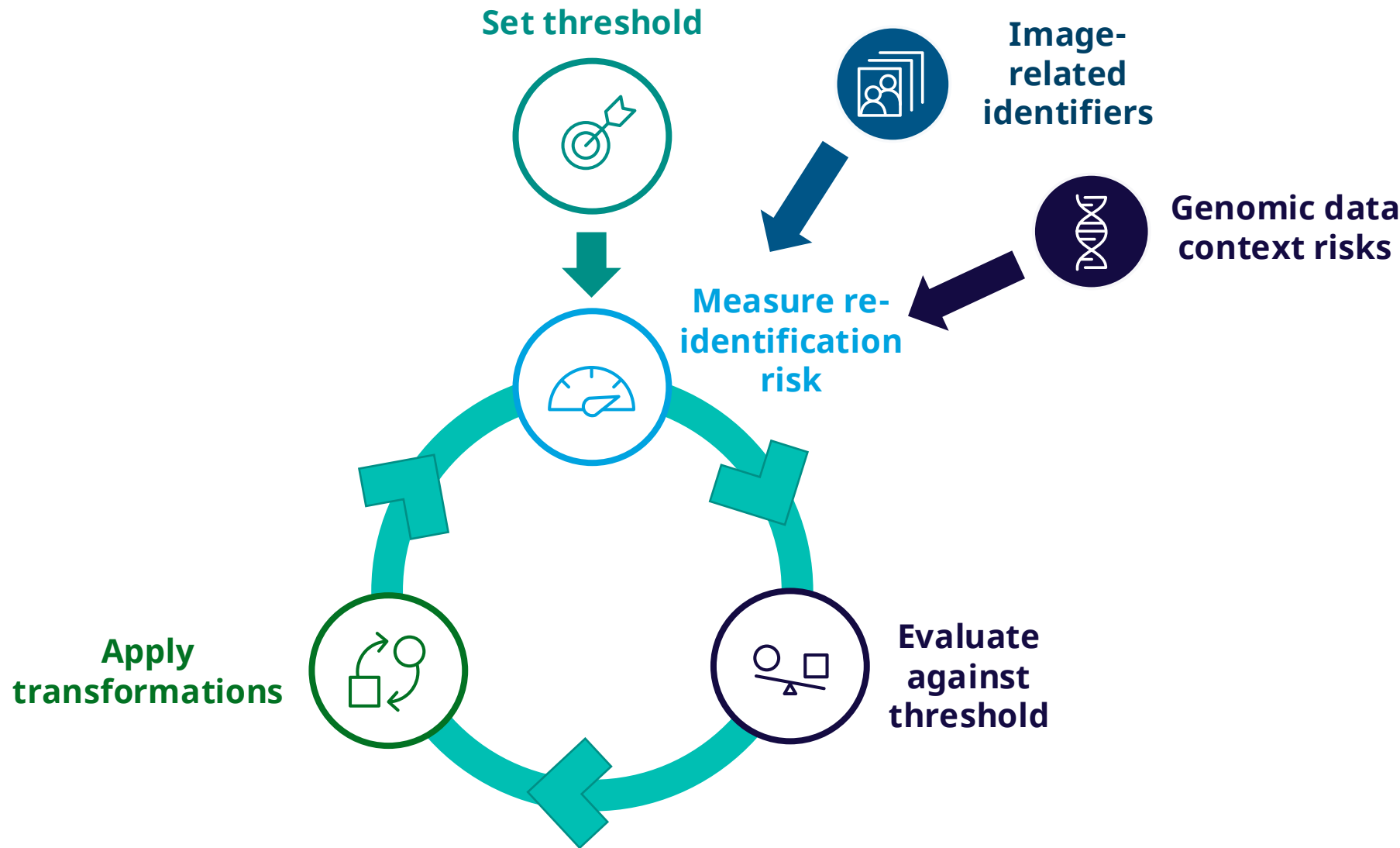
Statistical Anonymization Cycle



Statistical Anonymization Cycle

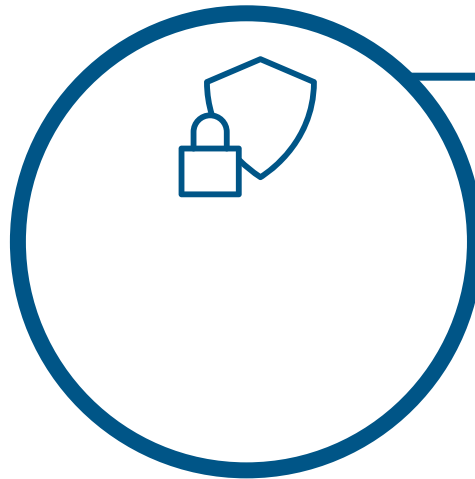


Statistical Anonymization Cycle



Global governance for responsible data re-use

Scalable multi-modal sharing depends on controlled access + auditability.



Core (foundation): Privacy and compliance

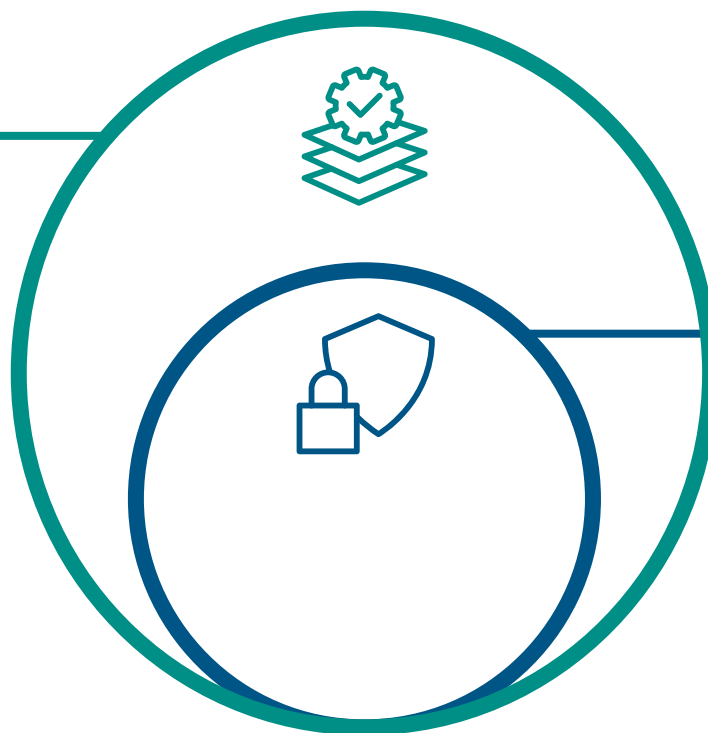
- Applicable privacy principles and laws
- Modality-specific privacy considerations
- Contextual risk across linked data
- Ethical data use & proportionate protections

Global governance for responsible data re-use

Scalable multi-modal sharing depends on controlled access + auditability.

Governance mechanisms

- Secure processing environments (SPEs)
- Access controls & permissions
- Consent & opt-out management
- Metadata cataloguing + audit trails

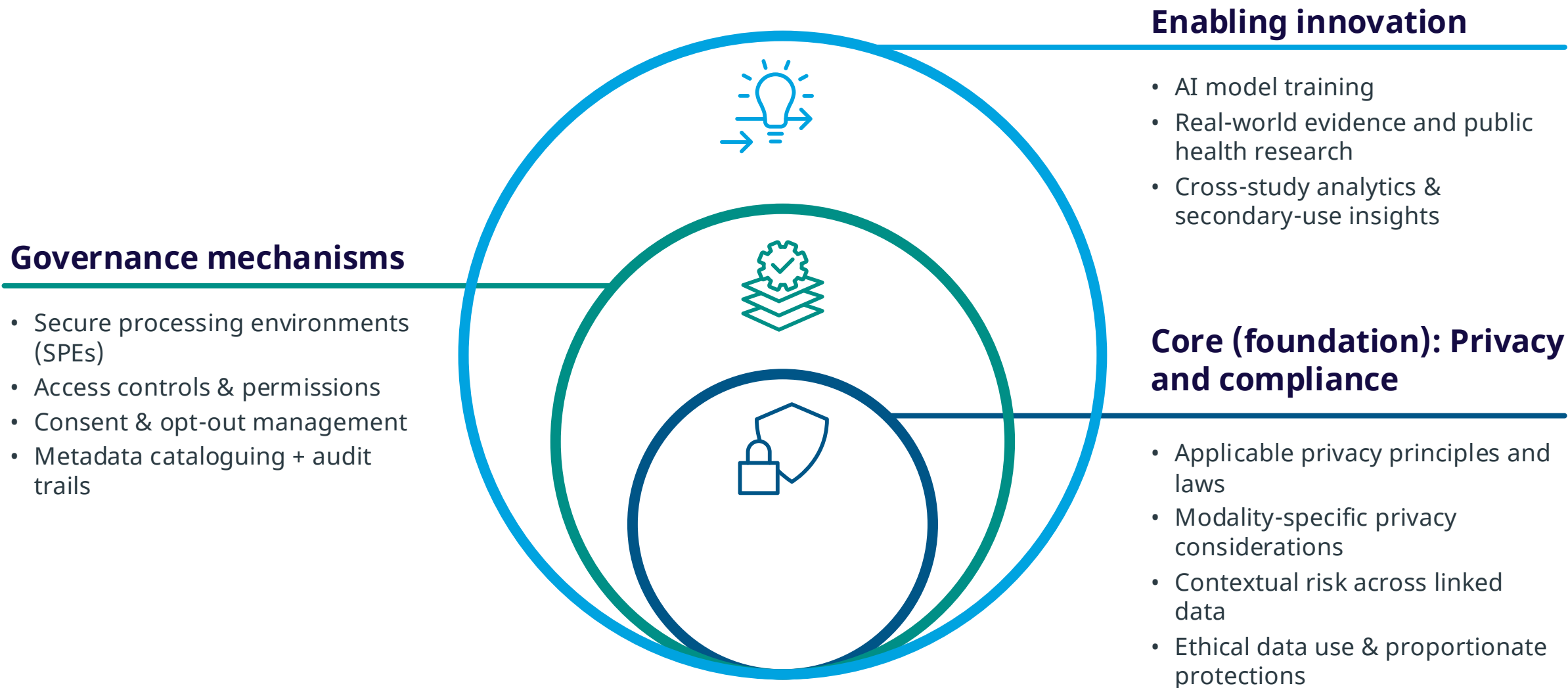


Core (foundation): Privacy and compliance

- Applicable privacy principles and laws
- Modality-specific privacy considerations
- Contextual risk across linked data
- Ethical data use & proportionate protections

Global governance for responsible data re-use

Scalable multi-modal sharing depends on controlled access + auditability.



Multi-modal anonymization for external data sharing

Situation

- A clinical trial sponsor aimed to share:
 - > **Structured datasets**
 - > **DICOM images**
 - > **Genomic data**
- All data was to be released under a **controlled-access platform**, with genomic data shared unaltered alongside anonymized structured and imaging data.

Multi-modal anonymization for external data sharing

Situation

- A clinical trial sponsor aimed to share:
 - > **Structured datasets**
 - > **DICOM images**
 - > **Genomic data**
- All data was to be released under a **controlled-access platform**, with genomic data shared unaltered alongside anonymized structured and imaging data.

Challenge

- Apply a unified risk strategy across structured and imaging modalities
- Assess privacy risks of releasing non-anonymized genomic data
- Identify potential linkage attacks across modalities
- Recommend mitigations without altering genomic data

Multi-modal anonymization for external data sharing

Situation

- A clinical trial sponsor aimed to share:
 - > **Structured datasets**
 - > **DICOM images**
 - > **Genomic data**
- All data was to be released under a **controlled-access platform**, with genomic data shared unaltered alongside anonymized structured and imaging data.

Challenge

- Apply a unified risk strategy across structured and imaging modalities
- Assess privacy risks of releasing non-anonymized genomic data
- Identify potential linkage attacks across modalities
- Recommend mitigations without altering genomic data

Solution

- Conducted joint risk quantification across structured and imaging data
- Applied harmonized anonymization to preserve utility and privacy
- Performed a three-part genomic risk consultation:
 - > Literature review of theoretical privacy threats
 - > Review of sponsor's data protection practices
 - > Interviews with SMEs to assess data sensitivity

Multi-modal anonymization for external data sharing

Situation

- A clinical trial sponsor aimed to share:
 - > **Structured datasets**
 - > **DICOM images**
 - > **Genomic data**
- All data was to be released under a **controlled-access platform**, with genomic data shared unaltered alongside anonymized structured and imaging data.

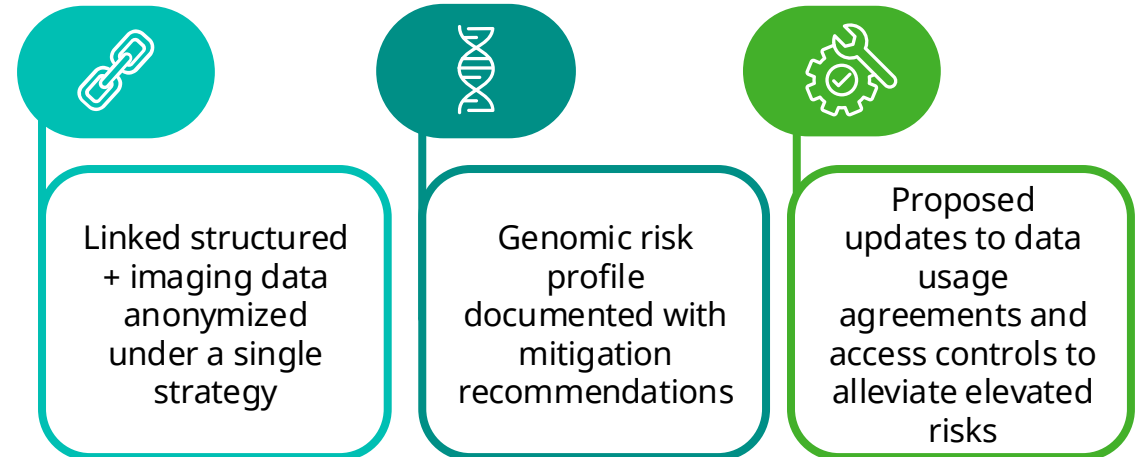
Challenge

- Apply a unified risk strategy across structured and imaging modalities
- Assess privacy risks of releasing non-anonymized genomic data
- Identify potential linkage attacks across modalities
- Recommend mitigations without altering genomic data

Solution

- Conducted joint risk quantification across structured and imaging data
- Applied harmonized anonymization to preserve utility and privacy
- Performed a three-part genomic risk consultation:
 - > Literature review of theoretical privacy threats
 - > Review of sponsor's data protection practices
 - > Interviews with SMEs to assess data sensitivity

Results





Final Consideration

The future of clinical transparency will be shaped by our ability to responsibly unlock the value of multimodal data. By aligning processes early and strengthening governance across imaging, wearables and genomics, we set the foundation for secure data sharing that supports scientific advancement and patient trust.



Thank you!