

NATIONAL
ACADEMIES

Sciences
Engineering
Medicine

Toward a 21st Century National Data Infrastructure

Managing Privacy and Confidentiality Risks
with Blended Data

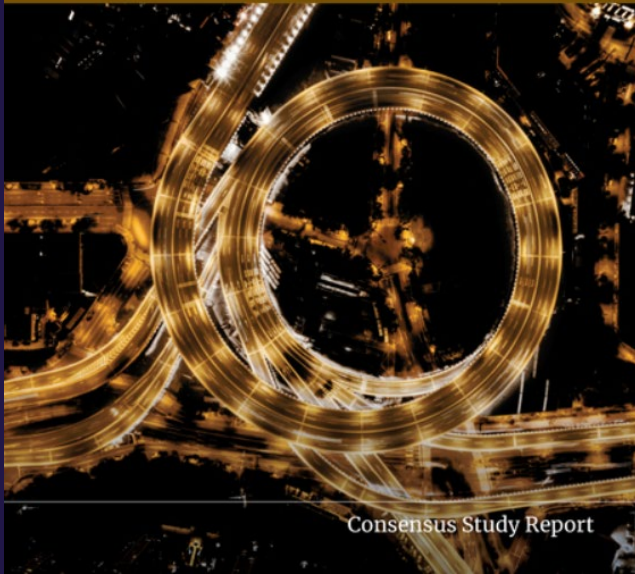
CNSTAT Spring Seminar 4/1/24

The Committee on National Statistics

The Division of Behavioral and Social Sciences and Education



Toward a 21st Century
National Data Infrastructure:
Managing Privacy and Confidentiality
Risks with Blended Data



Consensus Study Report

*Toward a 21st Century National
Data Infrastructure: Managing
Privacy and Confidentiality
Risks with Blended Data*

<https://nap.nationalacademies.org/catalog/27335/toward-a-21st-century-national-data-infrastructure-managing-privacy-and-confidentiality-risks-with-blended-data>

Panel Members

- **Jerry Reiter**, Duke University
- **Claire Bowen**, Urban Institute
- **Aloni Cohen**, University of Chicago
- **Diana Farrell**, National Bureau of Economic Research
- **Robert Goerge**, University of Chicago
- **Nick Hart**, Data Foundation
- **Hosagrahar Jagadish**, University of Michigan
- **Dan Kifer**, The Pennsylvania State University
- **Karen Levy**, Cornell University
- **Salomé Viljoen**, Michigan Law School
- **Mark Watson**, Federal Reserve Bank of Kansas City (formerly)

Structure of the Report

1. Introduction
2. Technical Approaches to Managing Risk When Sharing Blended Data
3. Policy Approaches to Managing Risk When Sharing Blended Data
4. A Model Framework for Decision Making When Sharing Blended Data
5. Conclusions

Issues In Managing Blended Data Risk:

Risk Spans the Blended Data Life Cycle

- The blended data life cycle spans:
 - initial conceptualization of blended data;
 - identifying and accessing ingredient data sources;
 - blending the data from those sources; and
 - sharing the resulting data products.
- Each of these stages presents potential risks to privacy and confidentiality, and subsequent harms to data subjects and data holders.
- Disclosure risks and harms can be magnified in blended data.

Issues In Managing Blended Data Risk:

Risks In Blended Data Can Be Managed

- No non-trivial data release method guarantees zero risks to privacy. Generally, providing greater access enhances usefulness, but also increases disclosure risks for data subjects.
- As a general rule, enhancing the usefulness of blended data requires accepting greater disclosure risks.

Trade-offs in disclosure risks, disclosure harms, and data usefulness are unavoidable and are central considerations when planning data-release strategies, particularly for blended data. Effective technical approaches to manage disclosure risks prioritize the usefulness of some analyses over others. (Conclusion 2-1)

Tools for Managing Risk in Blended Data

- The report describes the potential and limitations of existing technical approaches including:
 - Secure multiparty computation
 - Synthetic data with validation/verification
 - Classical statistical disclosure limitation
 - Formal privacy
- Policy approaches (e.g., laws, regulations, data enclaves and licenses) are essential components of the life cycle of blended data. They describe relationships of trust in data use.

Key Attributes of a Framework for Managing Risk

- Determining appropriate access depends on the acceptability of the privacy risk, given anticipated usefulness and potential harm.
- A framework that accounts for the unique attributes of blended data can assist decision-making.

Key Attributes of a Framework for Managing Risk:

Responds to Stakeholder Interests

- Engagement with stakeholders, including data holders, data users, and decision makers, is important for effective management of trade-offs.
- Ideally, this occurs throughout the design and implementation of privacy-protection strategies.
- Communication plans may differ depending on the needs of relevant groups:
 - For the public, use plain language to describe context-specific protections.
 - For data users, include methods for demonstrating data quality after privacy protections are applied.

Effective communication with data holders and data users can help agencies understand and better manage disclosure risk/usefulness trade-offs. (Conclusion 2–2)

Key Attributes of a Framework for Managing Risk:

Adapts to Policy and Technology Changes

- As policy priorities change, data availability can change. As more data are made available, the potential for privacy risk also increase. Technical approaches to limit privacy risk are advancing.
- Even when regulatory guidance and procedures for managing privacy risks are established, social acceptance of sharing and use of blended data will change.

The effectiveness of a framework for making decisions about acceptable disclosure risks given expected usefulness of data depends on whether that framework is dynamic. A dynamic framework allows for changing policy needs and data availability over time, in a way that accounts for the interests of data subjects, data holders, and data users. (Conclusion 3–1)

Key Attributes of a Framework for Managing Risk:

Reflects Different Levels of Risk and Usefulness

- Acceptable disclosure risk is a policy decision.
- As uses and users of blended data may have differing needs, policy can establish tiered access, describing levels of potential risk, harm, and usefulness and procedures in place to secure data access.

Tiered access for data users and agencies is a key component of a dynamic disclosure risk/usefulness framework, to reflect differences in acceptable risks given policy priorities. (Conclusion 3–2)

Key Attributes of a Framework for Managing Risk:

Provides a Common Lexicon For Communication

- Coordinating best practices for risk management across data holders and data users across disciplines requires a shared language reflecting the concepts of risk, harm, and usefulness.
- Shared language also enables quantification of these concepts, enabling them to be considered when managing trade-offs.

A common, cross-disciplinary language and lexicon describing privacy and confidentiality risks and harms, as well as data usefulness, is needed. Interpretable and measurable terms can promote meaningful discussions among stakeholders, including data subjects and decision makers. (Conclusion 3–3)

Other Aspects of a Framework for Managing Risk:

Documents Calculations, Assumptions, and Decisions

- A model to manage risk in blended data supports transparency.
- Documenting the process of preparing and analyzing blended data supports understanding the potential benefits and limitations of blended data products.
- Documenting assessments of disclosure risk and anticipated usefulness, assumptions of potential harms, and decisions regarding acceptable risk can inform communication with stakeholders, and inform future decision making, including justification or revision of approaches.

A Model Framework for Managing Risk:

Six Steps Supported with Guiding Questions

- Drawing from the panel's review of technical and policy approaches, the panel provides a framework that accounts for the attributes of blended data for making decisions about data-protection methods.
- Framework encourages agencies to answer a set of questions at each stage of the data-blending lifecycle to aid decision-making. Rather than attempting to cover all data-blending scenarios or stipulate precise approaches, the framework provides a lens to promote careful consideration of key questions.

Technical and policy approaches in combination are necessary for effective management of disclosure risks. (Conclusion 4–1)

A Model Framework for Managing Risk:

1. Determine auspice and purpose of the project

- a. What are the anticipated final products of data blending?
- b. What are potential downstream uses of blended data?
- c. What are potential considerations for disclosure risks and harms, and data usefulness?

A Model Framework for Managing Risk:

2. Determine ingredient data files

- a. What data sources are available to accomplish blending, and what are the interests of data holders?
- b. What steps can be taken to reduce disclosure risks and enhance usefulness when compiling ingredient files?

A Model Framework for Managing Risk:

3. Obtain access to ingredient data files

- a. What are the disclosure risks associated with procuring ingredient data?
- b. What are the disclosure risk/usefulness trade-offs in the plan for accessing ingredient files?

A Model Framework for Managing Risk:

4. Blend ingredient data files

- a. When blending requires linking records from ingredient files, what linkage strategies can be used?
- b. Are resultant blended data sufficiently useful to meet the blending objective?

A Model Framework for Managing Risk:

5. Select approaches that meet the end objective of blending

- a. What are the best-available scientific methods for disclosure limitation to accomplish the blended data objective, and are sufficient resources available to implement those methods?
- b. How can stakeholders be engaged in the decision-making process?
- c. What is the mitigation plan for confidentiality breaches?

A Model Framework for Managing Risk:

6. Develop and execute a maintenance plan

- a. How will agencies track data provenance and update files when beneficial?
- b. What is the decision-making process for continuing access to or sunsetting the blended data product, and how do participating agencies contribute to those decisions?
- c. How will agencies communicate decisions about disclosure management policies with stakeholders?

Figure 4-1. Model decision matrix of disclosure-protection strategies given potential harms and usefulness.

Potential Usefulness from Access (potential usefulness decreases from top to bottom)	Potential Harms Resulting from Disclosure (potential harm increases from left to right)				
	Negligible	Minor and fleeting	Significant and lasting	Life altering	Life threatening
Assess policy—major impact	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: minimal 	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: light 	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: rigorous 	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: rigorous 	<ul style="list-style-type: none"> • <u>Access</u>: RD • <u>Privacy</u>: very rigorous
Assess policy—modest impact	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers; DL • <u>Privacy</u>: minimal 	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: light 	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: rigorous 	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: rigorous 	<ul style="list-style-type: none"> • <u>Access</u>: RD; tiers • <u>Privacy</u>: very rigorous
General knowledge	<ul style="list-style-type: none"> • <u>Access</u>: tiers; DL • <u>Privacy</u>: minimal 	<ul style="list-style-type: none"> • <u>Access</u>: tiers; DL • <u>Privacy</u>: light 	<ul style="list-style-type: none"> • <u>Access</u>: tiers; DL • <u>Privacy</u>: rigorous 	<ul style="list-style-type: none"> • <u>Access</u>: tiers; DL • <u>Privacy</u>: very rigorous 	<ul style="list-style-type: none"> • <u>Access</u>: DL • <u>Privacy</u>: very rigorous
Educational	<ul style="list-style-type: none"> • <u>Access</u>: DL • <u>Privacy</u>: minimal 	<ul style="list-style-type: none"> • <u>Access</u>: DL • <u>Privacy</u>: rigorous 	<ul style="list-style-type: none"> • <u>Access</u>: DL • <u>Privacy</u>: rigorous 	Do not blend data	Do not blend data

Case Studies

- **Case Study A: Blending Federal Data**

At the workshop, Dauberman and Arnesberger described blending ingredient data from the Department of Education's College Scorecard and Statistics of Income data from the Internal Revenue Service.

- **Case Study B: Blending Federal and State Data**

At the workshop, Penner described blending student-level data from state educational agencies with earnings data from IRS, which are housed at the Census Bureau for use on approved projects.

- **Case Study C: Blending State Data**

The Multi-State Postsecondary Report blends education and earnings data from multiple state agencies through a Multi-State Data Collaborative maintained by the Coleridge Initiative.

Summary of Conclusions (1)

- Agencies, policymakers, data users, and data subjects need to recognize that any blended (or nonblended) data release that offers nontrivial usefulness introduces disclosure risks; it is not productive or correct to think of disclosure risks as a “yes or no” feature.
- Data-release strategies need to balance disclosure risks with data usefulness. When usefulness is high, stakeholders may be willing to accept greater risks to realize the benefits. Agencies can use various disclosure-protection methods for differing data-analysis objectives, such as tiered access approaches.

Summary of Conclusions (2)

- Successful risk-management strategies are likely to involve both technical and policy approaches. Some existing approaches can be gainfully applied with blended data, but others are less effective given the magnified disclosure risks in blended data.
- Disclosure risk management approaches need to be dynamic, involve stakeholder input, and rely on best practices. These characteristics can help determine desirable disclosure risk/usefulness trade-offs.
- Agencies can be (and should be, in the panel's opinion) intentional in examinations of risks at all stages of the blended data lifecycle.

Areas for Further Study

- Some key issues called out in report:
 - accounting for effects of composition in practice
 - obtaining stakeholder feedback
 - communicating disclosure risk/usefulness trade-offs to the public
- Two areas for future studies:
 - informed consent
 - research computing and data workforce