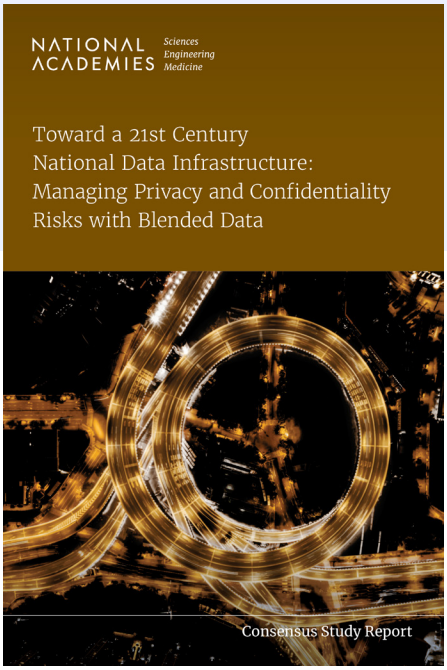
Toward a 21st Century National Data Infrastructure



Managing Privacy and Confidentiality Risks with Blended Data

Significant technical advances and policy changes have increased the availability of data that can be used to inform evidence building. *Blended data*—combined sources of previously collected data—can improve the quality of analyses, enable new analyses, and reduce burden and cost to the public. Recent federal legislation, regulation and guidance has described broadly the roles and responsibilities for stewardship of blended data. Yet, questions remain as the country strives to create a modern national data infrastructure.

THE DATA INFRASTRUCTURE SERIES

To answer those questions, the National Academies of Sciences, Engineering, and Medicine, with funding from the National Science Foundation, appointed three consensus study panels to develop a vision for a new data infrastructure for national statistics and social and economic research. Each of these reports[1] was prepared by its own diverse panel with expertise well suited to its particular focus, and each was informed by public workshops facilitating the exchange of experiences among academics, federal agencies, and civil society.

THIS REPORT

For this third report, the panel identified privacy and confidentiality aspects of sharing and analyzing blended data:

- that can be addressed by technical approaches;
- those that may require policy approaches; and

[1] See https://www.nationalacademies.org/our-work/toward-a-vision-for-a-new-data-infrastructure-for-federal-statistics-and-social-and-economic-research-in-the-21st-century.

- key attributes of a framework to guide best practices when designing and evaluating integrated technical and policy approaches.

Taking these findings into account, the panel provided a framework illustrated using case studies informed by the panel's public workshop event.

ISSUES IN MANAGING BLENDED DATA RISK

Risk Spans the Blended Data Life Cycle

The blended data life cycle spans the 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.

Blended Data Magnifies Disclosure Risk...

Data linkage often requires data holders to share identification numbers, names, or other confidential fields. In addition, blended data often have multiple data owners. Because large parts of data records potentially are available to many individuals, there may be greater opportunity for ill-intentioned users to learn confidential information from blended data shared without adequate disclosure protection. What may seem like a safe release strategy may be undone by other data holders' actions.

...And Subsequent Harms

Disclosure harms also may be magnified, particularly when data blending is used to add sensitive variables (e.g., education, economic, health), which are needed to inform policy.

Acceptable Risk Is a Policy Decision

Data privacy and access statutes, regulations, and policies determine the persons, places, and purposes in which protected data may be used. This is informed by an assessment of acceptable risk given anticipated usefulness.

Risks in Blended Data Can Be Managed

No non-trivial data release method guarantees zero risks to privacy and confidentiality. Providing greater access to the blended data enhances data usefulness, but it 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)

Better Measurement Can Assist Decisions

Measurement of disclosure risk, disclosure harm, and data usefulness provides a structured way to assess trade-offs when releasing protected data products. Measurement is particularly challenging for blended data, though there are effective approaches in the research literature applied by some agencies today.

TOOLS FOR MANAGING RISK

Data holders have a variety of technical and policy tools to manage these risks.

Technical Approaches Can Be Applied Throughout the Blended Data Life Cycle

Methods are available like secure multiparty computation, synthetic data, and differential privacy that offer ways to reduce risks. Some technical controls are deployable now for a given context and scale, but others require more research.

Apply Technical Controls with Policy Controls

Some disclosure risks cannot be mitigated with technical controls alone. Policy controls, such as laws, regulations, and data enclaves and licenses, are an essential component of all stages of the life cycle of blended data. They describe relationships of trust in data use. Augmenting technical controls with targeted policies can manage these risks more effectively. Additionally, transparent processes can legitimate blended data uses that have moved outside of the original contexts in which ingredient data were provided.

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

KEY ATTRIBUTES OF A FRAMEWORK FOR MANAGING RISK IN BLENDED DATA

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

Responds to Stakeholder Interests

Engagement with stakeholders, including data holders, data users, and decision makers, is important for effective management of trade-offs. Engagement best occurs throughout the design and implementation of privacy- and confidentiality-protection strategies. Communication plans may differ depending on the needs of relevant groups. For the public, plans ideally use plain language to describe context-specific protections. For data users, plans are most helpful when they 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)

Adapts to Policy and Technology Changes

As policy priorities change, data availability can change. As more data are made available, the potential for disclosure risk also increases. Technical approaches to limit disclosure risk are advancing. Even when regulatory guidance and procedures for managing disclosure 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)

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)

Assists Coordination Among Decision Makers

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.

Provides a Common Lexicon for Effective Communication

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)

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 of 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

Drawing from the panel's review of technical and policy approaches and considerations, the panel provides a model framework for making decisions about data-

protection strategies that accounts for the unique attributes of blended data. The framework encourages agencies to answer a set of questions at each stage of the data-blending life cycle 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.

The report applies the framework to three case studies in the domain of education (blending federal data, federal and state data, and state data). In each, the discussion illustrates how the framework can be used to support considered decision-making.

A Framework for Managing Disclosure Risks in Blended Data

1. **Determine auspice and purpose of the blended data project.**
   a. What are the anticipated final products of data blending?
   b. What are the potential downstream uses of blended data?
   c. What are potential considerations for disclosure risks and harms and data usefulness?

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?

3. **Obtain access to the ingredient data files.**
   a. What are the disclosure risks associated with procuring ingredient data?
   b. What are the disclosure risk and usefulness trade-offs in the plan for accessing ingredient files?

4. **Blend the 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?

5. **Select approaches that meet the end objective of data 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?

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?

PANEL ON APPROACHES TO SHARING BLENDED DATA IN A 21ST CENTURY DATA INFRASTRUCTURE **Jerome P. Reiter** (*Chair*), Duke University; **Claire McKay Bowen,** Urban Institute; **Aloni Cohen,** University of Chicago; **Diana Farrell,** National Bureau of Economic Research; **Robert M. Goerge,** University of Chicago; **Nicholas Hart,** Data Foundation; **Hosagrahar V. Jagadish,** University of Michigan; **Daniel Kifer,** The Pennsylvania State University; **Karen Levy,** Cornell University; **Salomé Viljoen,** University of Michigan Law, Harvard University, and Cornell Tech; **Mark Watson,** Federal Reserve Bank of Kansas City (formerly)

STAFF **Jennifer Park,** Study Director; **Bradford Chaney,** Senior Program Officer; **Anthony Mann,** Senior Program Coordinator; **Kevona Jones,** Senior Program Assistant

FOR MORE INFORMATION

This Consensus Study Report Highlights was prepared by the Committee on National Statistics based on the Consensus Study Report *Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data* (2024).

The study was sponsored by the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project. Copies of the Consensus Study Report are available from the National Academies Press, (800) 624-6242 or https://www.nap.edu/catalog/27335.

Division of Behavioral and Social Sciences and Education

NATIONAL ACADEMIES *Sciences Engineering Medicine*