CASE STUDY #5: Minimizing Access to PII: Best Practices for Access Controls and Disclosure Avoidance Techniques

Illustrating best practices for minimizing access to sensitive information with education data maintained in a statewide longitudinal data system

About PTAC

The U.S. Department of Education established the Privacy Technical Assistance Center (PTAC) as a "one-stop" resource for education stakeholders to learn about data privacy, confidentiality, and security practices related to student-level longitudinal data systems and other uses of student data. PTAC provides timely information and updated guidance through a variety of resources, including training materials and opportunities to receive direct assistance with privacy, security, and confidentiality of student data systems. More PTAC information is available at https://studentprivacy.ed.gov.

PTAC welcomes input on this document and suggestions for future technical assistance resources relating to student privacy. Comments and suggestions can be sent to PrivacyTA@ed.gov.

Purpose

In December 2011, the U.S. Department of Education (Department or we) released new regulations governing the Family Educational Rights and Privacy Act (FERPA), (76 FR 75604 (Dec. 2, 2011)), and supplemental non-regulatory guidance. We are providing the following case study to illustrate how specific provisions of FERPA may be implemented. This case study uses fictional agencies, does not address individual circumstances, and does not consider additional legal requirements that may be required under other Federal, state, or local laws.

We will release additional case studies, and welcome suggestions for future topics. Comments and suggestions can be sent to PrivacyTA@ed.gov.

Background

The state education agency (SEA) in State X operates a statewide longitudinal data system (SLDS) that contains a large quantity of personally identifiable information (PII) from students' K-12 and postsecondary education records, which are protected from unauthorized disclosure by the Family Educational Rights and Privacy Act (FERPA). In addition to the data governance and security policies that the SEA has adopted to protect the data under FERPA (see Data Security Checklist), due to the quantity and sensitivity of these data, the SEA also wants to follow additional best practices for data minimization.[1] Consequently, the SEA decides to implement a variety of access controls and disclosure avoidance methods in order to minimize access to sensitive information within the SEA, and to protect against unauthorized disclosure of PII outside of the SEA.

[1] Data Minimization refers to the Fair Information Practice of "only collecting personally identifiable information that is directly relevant and necessary to accomplish the specified purpose(s). [And for] only retaining personally identifiable information for as long as is necessary to fulfill the specified purpose(s)." It also extends to only allowing access to specific PII elements to those individuals who have a legitimate need to view and utilize those elements. See the National Center for Education Statistics (NCES) Technical Brief #2 Data Stewardship: Managing Personally Identifiable Information in Electronic Student Education Records (NCES 2011-602).

What does the SEA do?

In order to minimize access to sensitive information within the SLDS, the SEA follows data minimization best practices and implements role-based access controls on all student-level information. SEA employees' levels of access are determined by their job functions and responsibilities, in accordance with State X's SLDS data governance plan, and are implemented through appropriate physical and information technology (IT) security controls. Because the data collected and maintained by the SEA are also made available to external researchers and published in a variety of public reports (see below), the data governance plan also establishes access controls and disclosure avoidance measures for external dissemination of the data. The levels of access and their corresponding data minimization procedures identified by the SEA are as follows:

- *Raw Individual Student Data (contains direct identifiers, including Social Security Numbers [SSNs])*

  Integrating students' records into the SLDS requires the use of a number of direct identifiers (typically student's name, address, parents' names, and student's SSN or other unique student ID number) to identify specific students' records in datasets from different sources, and to link those records together longitudinally. Because of the sensitivity of these direct identifiers (especially SSNs), the SEA restricts access to the Raw Individual Student Data that contain those identifiers to the individuals directly responsible for data integration and record linkage. In State X's SEA, the data team identifies the sensitive direct identifiers and performs the record manipulation and linkage. These individuals follow the best practice of using the direct identifiers exclusively for the administrative purposes of performing the linkages, integrating the data into the SLDS, and performing quality assurance on the resulting linked files. Once the linkages are complete, the most sensitive direct identifiers are removed from the file along with a copy of the student record system student identification (SID) numbers. This "crosswalk file," linking the direct indicators to the student record system SID numbers, is stored separately in a secure electronic environment for use with any further linkages (see the NCES Technical Brief #2).

- *Redacted Individual Student Data (direct identifiers have been removed)*

  Redacting the direct identifiers reduces the overall sensitivity of the file. However, the redacted data file still contains PII, in the form of indirect identifiers (e.g., date of birth) and other identifying characteristics (e.g., race, gender, and disability status), data on education program participation, and on the student's teacher(s) that could be used to re-identify specific individuals. Consequently, the data are still protected by FERPA. Most of the statistical analysis performed by the SEA's employees is done using this redacted file.

  Periodically, the SEA receives requests from external data users (e.g., faculty researchers at the local university) interested in using student-level data to evaluate education policy questions. While most research can be performed using the publicly available de-identified individual student data (see below), there are times when the de-identified data may not have sufficient detail or precision for advanced analyses. In these cases, the SEA may enter into a written agreement with the external researchers to designate them as the SEA's "authorized representative," and provide the researchers with access to FERPA-protected student-level data maintained by the SEA for evaluations of Federally- or state-supported education programs (see the PTAC Case Study #4). When evaluating these requests from external researchers, the SEA adopts best practices for minimizing access to PII and evaluates the requestors' specific needs for direct identifiers. In many cases, these requestors' research needs can be met by using the redacted individual student data, without requiring the access to the more sensitive raw individual student data.

- *Aggregate Data Tables* (*the need to protect small cells*)

  To meet legal requirements, the SEA must publish various school and student performance indicators in aggregate tables. For example, the SEA uses data in the SLDS to construct aggregate data tables of student achievement broken down by various subgroups. Because many of these aggregate data tables contain information for small subgroups, the tables contain numerous cells with only one or two students in them. Consequently, these tables still contain PII, because it may be possible to identify specific individuals within those small cells based on one or more uncommon characteristics. In order to comply with the privacy requirements of FERPA, as well as the confidentiality and privacy provisions in both the Individuals with Disabilities Education Act (IDEA) and the Elementary and Secondary Education Act (ESEA), the SEA restricts access to these aggregate data tables (i.e., does not publish them) until sufficient disclosure avoidance measures have been taken to mitigate the risk of re-identification (see below).

- *Public Aggregate Data Tables* (*disclosure avoidance measures have been applied*)

  In order to release the aggregate data tables to the public, the SEA must perform disclosure avoidance analyses on the tables to identify potential disclosures, and then apply disclosure avoidance techniques to mitigate the risk that a reasonable person in the school community could identify specific students within the small cells of the tables. In this case, the SEA in State X decides to accomplish this by utilizing a disclosure avoidance technique known as "complementary cell suppression," whereby all cells in the table that fall below a particular threshold chosen by SEA (in this case n<5) are suppressed. The SEA then suppresses a select number of additional cells to prevent the possibility that the suppressed small cells could be re-calculated by subtracting the other reported cells from the tables' row and column totals. Once these suppressions have been applied (and audited), the SEA confirms that the tables no longer contain any PII, and the tables are published on the SEA's website.

- *De-identified Individual Student Data* (*disclosure avoidance measures have been applied*)

  After publishing the public aggregate data tables on its website, the SEA receives a number of requests from researchers and advocacy groups requesting additional data. These requestors explain that the public tables indicate that there may be some interesting trends in the data, and that they want to perform more extensive analyses on the student-level data. Recognizing the potential public value of these evaluations, the SEA decides to create a public-use version of the file. To accomplish this, the SEA takes the redacted individual student data file and removes or blurs any remaining indirect identifiers (e.g., replacing date of birth with year of birth). To de-identify the data further, the SEA then applies additional disclosure avoidance on the data, in this case by performing a perturbation technique, such as "swapping" (in which a statistical algorithm is used to swap data elements for a small number of individuals). At this point, to verify that the data have been properly de-identified, the SEA decides to have a statistician analyze the resulting file. The SEA's data owner makes the determination that the data cannot be used to re-identify any individual in the file with any reasonable certainty, thus satisfying the de-identification standards of FERPA (34 CFR § 99.31(b)) and making the data suitable for public release. The SEA then publishes the de-identified, disclosure-protected individual student data file on its website for public use.

What is the SEA permitted to do under FERPA?

FERPA permits the SEA's employees and authorized representatives to access PII from education records to audit or evaluate Federally- or state-supported education programs, (34 CFR 99.31(a)(3) and 99.35), and requires that all PII from education records be adequately protected from inadvertent or unauthorized re-disclosure and destroyed when no longer needed for the purposes of the evaluation (34 CFR 99.35). Using the FERPA requirements as a minimum, it is then a widely accepted best practice for SEAs to adopt broad data minimization practices and to apply additional restrictions and protections to those data, files, or systems containing PII elements generally considered to have higher potential for harm or misuse, like SSN and other direct identifiers.

Additional Resources

The resources below include links to federal regulations and several guidance documents providing more in-depth discussion of recommendations and techniques that can be used to de-identify tabular as well as student-level data. While these recommendations may not be appropriate for every situation, they may provide a better understanding of the relevant concepts and issues involved in selecting and applying data de-identification methods to education data.

- Family Policy Compliance Office (FPCO), U.S. Department of Education: https://studentprivacy.ed.gov
- FERPA Regulations, U.S. Department of Education: www.ed.gov/policy/gen/reg/ferpa
- Federal Regulations Resources, U.S. Department of Education: www.ed.gov/policy/gen/reg/edpicks.jhtml
- *FERPA Regulations Amendment*. U.S. Department of Education (December 2, 2011): www.gpo.gov/fdsys/pkg/FR-2011-12-02/pdf/2011-30683.pdf
- *FERPA Notice of Proposed Rulemaking*. U.S. Department of Education (March 24, 2008): www.ed.gov/legislation/FedRegister/proprule/2008-1/032408a.html
- *FERPA Regulations Amendment*. U.S. Department of Education (December 9, 2008): www.ed.gov/legislation/FedRegister/finrule/2008-4/120908a.pdf
- Privacy Technical Assistance Center (2012): *Case Study #4: PTAC Technical Assistance*, available at https://studentprivacy.ed.gov/resources/case-study-4-ptac-technical-assistance.
- Privacy Technical Assistance Center (2012): *Data De-identification: An Overview of Basic Terms*, available at https://studentprivacy.ed.gov/resources/data-deidentification-overview-basic-terms.
- Privacy Technical Assistance Center (2012): *Frequently Asked Questions—Disclosure Avoidance*, available at https://studentprivacy.ed.gov/resources/frequently-asked-questions-disclsoure-avoidance.
- Privacy Technical Assistance Center (2011): *Data Governance and Stewardship, available at* http://ptac.ed.gov/sites/default/files/issue-brief-data-governance-and-stewardship.pdf.
- Privacy Technical Assistance Center (2011): *Data Security Checklist*, available at https://studentprivacy.ed.gov/resources/data-security-checklist.
- Privacy Technical Assistance Center (PTAC), U.S. Department of Education: https://studentprivacy.ed.gov
- U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics (2011): *SLDS Technical Brief 1: Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records* (NCES 2011-601), available at http://nces.ed.gov/pubs2011/2011601.pdf.
- U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics (2011): *SLDS Technical Brief 2: Data Stewardship: Managing Personally Identifiable Information in Electronic Student Education Records* (NCES 2011-602), available at http://nces.ed.gov/pubs2011/2011602.pdf.

- U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics (2011): *SLDS Technical Brief 3: Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting* (NCES 2011-603), available at http://nces.ed.gov/pubs2011/2011603.pdf.
- U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics (2011): *Technical Brief: Statistical Methods for Protecting Personally Identifiable Information in the Disclosure of Graduation Rates of First-Time, Full-Time Degree- or Certificate-Seeking Undergraduate Students by 2-Year Degree-Granting Institutions of Higher Education* (NCES 2012-151), available at http://nces.ed.gov/pubs2012/2012151.pdf.