

De-Identification and Education Meeting Notes

Friday, October 13, 2017

11:30am-1 pm

Topic: Deidentification and Education

86 Attendees

Michael Hawes, Director of Student Privacy Policy, U.S. Dept of Education

- What is PII?
 - More than the obvious information such as the name and SSN. It includes indirect identifiers that may be combined to identify the actual person with any reasonable certainty. Just about any information about a student may be identifiable depending on the context.
- FERPA's Confidentiality Standard:
 - They often use the rule of 3; if you can't separate one student from two other similarly situated students, then you probably have revealed too much.
- FERPA does not have a safe harbor provision similar to HIPAA's "Safe Harbor" provision.
- Aggregate data can still contain PII data. This presentation is about avoiding/mitigating risk to unauthorized disclosure.
- States have their own methods for reporting data. Many states don't report data on subgroups smaller than 10 but in general, state methods vary largely. The Privacy Technical Assistance Center is available to assist. The Help Desk email address is PrivacyTA@ed.gov
- Common Mistakes in Public Reporting:
 - States establish minimum n-size rules without taking into account the total population size. So if your population is 6, but your minimum n-size rule is 5, then states may proceed without caution; nevertheless, the population is relatively small in comparison to the n-size that identification is impossible.
 - When divulging statistics and establishing fixed top and bottom thresholds, take into account extreme value reporting. A better way to handle extreme values is to vary the thresholds you use to report numbers.
 - Reporting numbers with missing values, but still reporting the total number of students surveyed. This makes it easy for anyone else to do simple math to determine the suppressed values.
 - Small cells/values aren't the only problems. Large values can be a problem if they are extreme or can be used to calculate the values of the protected cells elsewhere.
- You must about all of the different ways you're reporting data and all of the different levels that are reporting information like subgroup, grade, school, district, states, and national. Therefore, if a cell is suppressed at one level, it needs to be suppressed in at

least one other reporting entity at the next level of aggregation. Furthermore, you must take into consideration how the entities around you report information related to your school. All of the different releases of the same data influence that data reportability.

- What Reporting Options Do You Have?
 - Suppression: removing data to prevent the identification of individuals in small cells or with unique characteristics. Suppressions may result in very little data being produced for small populations, and there is still a risk of disclosure since various entities are reporting on the same population.
 - Blurring: any method that reduces the precision of data that is presented to reduce the certainty of identification. This includes various types of aggregation or reporting your results as percentages. This may result in users not being able to make inferences about small changes in the data and reduce the ability to perform time-series or cross-case analysis.
 - Perturbation: making small changes to the data to prevent identification of individuals from unique or rare characteristics. This method can minimize loss of utility compared to other methods but it may be inappropriate for program data because it reduces the transparency and credibility of the underlying data. In other words, you can't know for sure what one school's graduation rate is since it tweaks the underlying data. This data may be easy to reverse engineer if they are given access to the original data.
- Which Should You Choose?
 - First you should ask, "How will these data be used?" so that you maximize the use of the data
 - You can adopt hybrids or multiple methods
 - If using suppression, be aware of the row/column totals, and related tables as complementary suppression will most likely be necessary to prevent others from recalculating the totals
 - Don't rely on suppressing underlying population/subgroup counts as they are often known or can be inferred
 - When reporting data in percentage, round to whole number whenever possible
 - Be careful when you report individual-level data as you may need some amount of perturbation
 - Be sure to audit your results. Essentially look at the result and see if you can reverse engineer that data to determine the true numbers.

**Adrian L. Peoples, Ed.D, Lead Education Associate, Data Management and Governance,
Delaware Department of Education, adrian.peoples@doe.k12.de.us, 302-735-4161**

- The Delaware Dept. of Education collects various types of data with many identifiers. The department creates a Master Person Index and then maps it to the unique identifiers using a highly complex algorithm. They are currently at a 90% match rate.
- The Dept. of Education are doing Pseudonymization, which is creating a pseudonym for each student and then mapping it back to their unique identifiers.
 - This allows the department to connect the unique identifier with the student over a large period of time, essentially conduct research.
 - It also allows for interoperability amongst various datasets regarding the student (for example health data with education data)
 - The data collected over a long period time allow the department to create a holistic picture of the student and follow them long term
- Some challenges:
 - Match rates are less than 100% so they may end up looking at duplicate IDs
 - There are incomplete datasets with missing school information or demographic information
 - Re-identification is possible depending on the dataset
- The Delaware Department of Education uses algebra to suppress identifiable information. If the data is below the threshold then it gets suppressed. The end result may end up with data that is blurred, bottom-coded, and suppressed. They also aggregated small demographics which results in White and Non-White categories. Each dataset was sent to the U.S. Dept of Education to see if it could be reverse-engineered.
- Challenges to Disclosure Avoidance:
 - Saving (and Reusing) Non-Disclosive Data
 - Audience Cross-Contamination- Sometimes an audience member has access to the data and wants to see it reported in a different way often skewing any de-identification methods
 - Transparency Concerns: What are you hiding?

Question & Answer

Question: How does the “reasonable person in the school community” differ from a “reasonable person” standard on its own? Is it intended to be a higher standard?

Hawe’s Answer: It’s a higher standard because in the school community, a reasonable person in the school community, there is going to be some basic information available about students, and you will know who the students are in the subgroups. You will know contextual information about students. An example is discipline data such as expulsions. This data is often reported in the news or published in an article. You’ll be able to put all of the pieces together due to outside resources and your familiarity with the community. You’re combining reasonable methods of de-identification with information that’s commonly known to individuals in the schools community. The reasonable person component is a fairly loose standard as well. It’s a balancing act. At the end of the day,

we're talking about reasonable person. Can a reasonable person in the community re-identify a student? A reasonable person when it comes to re-identification is not a statistician.

Question: What are the responsibilities/standard for companies to de-identify the student data they hold? What about when they are de-identifying it or preparing it for reporting purposes on behalf of the school?

Answer: Hawe's response: FERPA applies to education agencies and institutions. It's the educational agency that determines what the appropriate threshold is. It's ultimately the school that's responsible for an unauthorized disclosure. Also, check with the third parties to make sure that you approve of their methods for de-identification. Also, make sure that whatever methods are used are satisfactory with your educational methods and under the law.

People's Response: Delaware doesn't govern conversations between LEAs and their vendors. LEAs have to follow the FERPA guidelines and follow state law regardless of who they work with. Delaware takes on FERPA responsibilities with whatever information they are given. When they get vendor data, sometimes they publish it in aggregate format to the public. As a general rule, we stay out of the conversation, other than to say, you must follow the law.

Question: What is the data de-identification requirement - if any - for a company to use to improve their own product/other internal uses?

Answer: Some of that will depend on how the entity received the information in the first place. Under the FERPA exception, the company essentially becomes a school official and they must be under the direct control of the school or district with regard to the use and maintenance of that data. The first question to ask is does the company's agreement with the school or district allow them to use data for product development? Under their agreement with the school, they may be required to de-identify the data. It's important for the company to make sure their methods are satisfactory and meets the standards for the school and states. Regarding metadata, please look at the 2014 Protecting Student Privacy While Using Online Services at <https://tech.ed.gov/wp-content/uploads/2014/09/Student-Privacy-and-Online-Educational-Services-February-2014.pdf>.

Question: Many teachers want to sign their students up for a free web service or an app. In order to protect their student's privacy they don't use the student's full name. Example: DanS or SmithD. Is this acceptable?

Answer: It depends. FERPA is context dependent. FERPA prohibits the disclosure of PII from a student's education record to a third party unless an exception applies. If a teacher is signing students up for a service, and it's completely anonymized and when the student is using that service they aren't providing any additional information, then that may be permissible as there is no disclosure of PII. If however the teacher provides the log in, but as part of the use, the student is taking tests and the program is maintaining

information that is related to the student's education record, then there may be an issue as you're approaching FERPA, even with an anonymized login.

Question: Can you go over some basic information on how to de-identify individual student data in order to protect their identity as they utilize online resources for educational purposes? What is best practices regarding de-identifying student data? What should educators do to make sure they are protecting student identities and student generated content?

Answer: First, ask what information is being provided and what's being collected? Is it being done under one of the FERPA exceptions? How is the school exercising control? Assess the risks. Select a method or methods that could be used to de-identify and re-identify. There is a variety of things you could do; there is no succinct answer.

People's response: First: Send as little information as possible to identify the student; Second: Know what data is being collected in the systems; and Third: ask if it the information will be reported.

Question: One of the major de-identification challenges for LEAs/SEAs is dealing with public disclosures from USED and trying to make sure the data the state has put out doesn't allow for re-identifiability when combined with a federal data report. How do you deal with that?

Answer: There is also a risk that one release could impact another. Just try to make sure that whatever reports are being released are meeting the reasonable person standard with regarding to the likelihood of de-identification.