CENTER FOR DATA INNOVATION

March 25, 2016

The White House
1600 Pennsylvania Avenue NW
Washington, DC 20500

Re: Comments on the Precision Medicine Initiative Draft Data Security Policy Principles and Framework

To Whom It May Concern:

On behalf of the Center for Data Innovation (datainnovation.org), we are pleased to submit these comments in response to the White House's request for comments on the Precision Medicine Initiative's Draft Data Security Policy Principles and Framework.[1]

The Center for Data Innovation is the leading think tank studying the intersection of data, technology, and public policy. With staff in Washington, DC, and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as predictive analytics, open data, cloud computing, and the Internet of Things. The Center is a nonprofit, nonpartisan research institute affiliated with the Information Technology and Innovation Foundation.

The Precision Medicine Initiative (PMI) is an ambitious and laudable effort to transform the medical community's understanding of human health. By making data about patient lifestyles, environments, and genomes available to medical researchers, the PMI will move health care away from a one-size-fits-all approach and support the development of personalized treatments for some of the most challenging and dangerous diseases.

We commend the White House for defining voluntary guidelines for responsibly managing PMI data in the Draft Data Security Policy Principles and Framework. However, the White House

[1] "Precision Medicine Initiative: Draft Data Security Policy Principles and Framework," The White House, Accessed March 24, 2016, https://www.whitehouse.gov/webform/precision-medicine-initiative-draft-data-security-policy-principles-and-framework.

incorrectly concluded in the draft that “de-identified data held by a PMI organization could still be matched to an individual.” De-identification, when done properly, is an effective and reliable method of protecting potentially sensitive PMI data, and it is a long-established technique at both the federal and state level to protect sensitive patient health information.

Please find our responses to the relevant questions in the attached document.

Sincerely,

Daniel Castro
Director
Center for Data Innovation
dcastro@datainnovation.org

Joshua New
Policy Analyst
Center for Data Innovation
jnew@datainnovation.org

DATA DE-IDENTIFICATION

The PMI will rely on large-scale sharing, analysis, and reuse of sensitive patient data, including genetic data, insurance claims, demographic information, and medical records, and thus participating organizations will require high levels of security measures to safeguard this data. The Draft Data Security Policy Principles and Framework acknowledges that de-identification—the removal of identifying information such as name, date of birth, or social security number from a data set—is a powerful tool for protecting sensitive patient data.[2] However, it also incorrectly states that increasing computing power and the potential to combine disparate data sets to extract new information make de-identification an unreliable security control and recommends that organizations participating in the PMI do not rely on de-identification alone. While recommending adoption of multiple security controls is prudent for implementing a "defense in depth" strategy, suggesting that de-ideidentification is not a reliable tool is incorrect, could discourage organizations participating in the PMI from adopting this useful technique to protect patient data, and could suggest that the uses of de-identification in other contexts is inappropriate. Moreover, organizations may instead adopt unnecessary, overly-restrictive, and more expensive methods to protect data that limit the effective use of PMI data.

While no information security method, including de-identification, is perfect, the risk of re-identification when data is properly de-identified has been greatly exaggerated by some commentators, and it would be a mistake to reverse existing federal policy based on these inaccuracies.[3] The notion that de-identification is an unreliable tool is commonly promulgated by commentators misinterpreting primary literature or failing to recognize that many instances of re-identification were only possible because the data was improperly de-identified in the first place.[4] For example, if a data set is properly de-identified in accordance with the Safe Harbor Standard defined by the Health Insurance Portability and Accountability Act (HIPAA) which requires the removal or modification of 17 specific data elements, only 0.04 percent of

[2] "Precision Medicine Initiative: Draft Data Security Policy Principles and Framework," The White House, February 25, 2016, https://www.whitehouse.gov/sites/whitehouse.gov/files/documents/PMI_Security_Principles_and_Framework_FINAL_022516.pdf.

[3] Ann Cavoukian and Daniel Castro, "Big Data and Innovation, Setting the Record Straight: De-identification *Does* Work," Information Technology and Innovation Foundation, June 16, 2014, http://www2.itif.org/2014-big-data-deidentification.pdf.

[4] Ibid.

individuals are “uniquely identifiable.”[5] It is important to note the significant difference between uniquely identifying an individual, which means recognizing a specific, discrete set of characteristics within a data set, and re-identifying an individual, which means gleaning a person’s actual identity—his or her name, birth date, and so on—from a data set.

The Draft Data Security Policy Principles and Framework should encourage organizations to find the optimal balance for securing data without sacrificing utility. If organizations participating in the PMI are cautioned against using de-identification, they may be inclined to use alternative methods more commonly regarded as “safe” but that could substantially reduce the value of PMI data. For example, aggregating data can reduce the risk of identifying specific individuals in a data set and still provide researchers with insight into larger-scale trends, such as health characteristics of large demographic groups. However, this sacrifices researchers’ ability to study granular data about specific individuals, which is necessary to develop personalized treatments or understand trends about minority groups. And if data sets are collapsed into only a few summary statistics, any conclusions based on this data will likely be too broad to be useful. Similarly, data minimization—when data is only collected, analyzed, and stored for narrow, pre-defined purposes and discarded immediately after it fulfills this purpose—may help safeguard sensitive data, but would be at odds with the goals of the PMI. Regardless of the application, analysis of more complete and granular data sets can often produce more nuanced and profound insights than analysis of smaller data sets. In the context of the PMI, a data set of 1,000 genomes is immeasurably more valuable to research than a data set of 10 genomes, as researchers can compare a larger variety of genetic markers for specific conditions, identify more factors that influence how conditions such as diabetes or autism present themselves, better study drug interactions, and more. Discarding data after just one of these analyses would prevent the data from ever being re-identified, but would also preclude any other opportunity for beneficial analysis.

In addition, the Draft Data Security Policy Principles and Framework should encourage organizations to properly de-identify data. Many of the examples in popular media of the risk of re-identification come from data that has not been properly de-identified. By suggesting that de-identification is ineffective, the Draft Data Security Policy Principles and Framework would in effect discourage organizations from demanding high standards for de-identification. This would be like a dentist telling her patients that brushing and flossing their teeth is not very useful since

[5] Simon Cohn, “Ltr to Sec'y 2007 Dec 21 Secondary Uses,” National Committee on Vital and Health Statistics, December 21, 2007, http://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/071221lt.pdf.

many people who brush and floss get cavities. Instead, of course, the dentist should be instructing patients on how to properly brush and floss so that they can avoid cavities.

The existing PMI Privacy and Trust Principles call for “protection against any intentional or unintentional…re-identification of PMI data,” and entities holding PMI data should be encouraged to use proper de-identification techniques to achieve this goal.[6] Just as the Draft Data Security Policy Principles and Framework recommends participating organizations adopt a risk-based approach for protecting PMI data, so too should it recommend de-identification as the best risk-mitigating tool organizations have at their disposal for preventing re-identification of PMI data.[7]

[6] “Precision Medicine Initiative: Data Security Policy Principles and Framework,” The White House, February 25, 2016, https://www.whitehouse.gov/sites/whitehouse.gov/files/documents/PMI_Security_Principles_and_Framework_FINAL_022516.pdf.
[7] Ibid.