From: Schaeuble, John
To: Rykaczewska, Agnieszka@CDII; Atifeh, Sussan@CDII; Hess, Catherine@CDPH
Subject: articles for distribution
Date: Thursday, December 19, 2024 6:11:44 PM
Attachments: BigHealthDataResearch.pdf
SuccessOfReIdentifications.docx

This Message Is From an External Sender
This message came from outside your organization.

Hi, Neszka, Sussan, and Catherine

Before following up on the full committee's motion for the subcommittee to make further changes in the document about IPA reviews, I did some research on topics that had been discussed at the December meeting, artificial intelligence and re-identification, and I found two articles that I think should be made available to both the subcommittee and the full committee.

Doerr, M., & Meeder, S. (2022). Big health data research and group harm: The scope of IRB review. *Ethics & Human Research*, *44*(4), 34-38. https://doi.org/10.1002/eahr.500130

I've attached the entire article, which is available as open source. A summary of it appears below.

Authors' abstract: Much of precision medicine is driven by big health data research—the analysis of massive datasets representing the complex web of genetic, behavioral, environmental, and other factors that impact human well-being. There are some who point to the Common Rule, the regulation governing federally funded human subjects research, as a regulatory panacea for all types of big health data research. But how well does the Common Rule fit the regulatory needs of this type of research? This article suggests that harms that may arise from artificial intelligence and machine-learning technologies used in big health data research—and the increased likelihood that this research will affect public policy—mean it is time to consider whether the current human research regulations prohibit comprehensive, ethical review of big health data research that may result in group harm.

Rocher, L., Hendrickx, J. M., & de Montjoye, Y. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*(3069). https://doi.org/10.1038/s41467-019-10933-3

I've attached the introduction from this article, which is summarized below. The full text is available as open source.

Authors' abstract: While rich medical, behavioral, and socio-demographic data are key to modern data-driven research, their collection and use raise legitimate privacy concerns. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address those concerns. We here propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.

John Schaeuble
jschaeuble@csus.edu