# Reindentification Risk from the Mosaic Effect

**Please provide any feedback regarding this scenario in the comment form below or by clicking here.**

An NIH research team collects and manages data for a Large National Survey wherein consented participants agree to allow linkage of their survey data to administrative healthcare claims and blood samples used for laboratory testing of biomarkers for chronic disease. Until 2014, the research team stewarding the data had two separate releases: a public release of an anonymized data set (Public Data Set) and a restricted release (Restricted Data Set) with the claims data and laboratory tests, but no direct identifiers. Participants consent to terms of release for both of these data sets, with the understanding that greater scrutiny and safeguards are required for researchers using the restricted data, and that the Public Data would be available online.

In 2013, NIH funded Max Researcher to use the Restricted Data Set conduct exploratory Precision Medicine. Max integrates the restricted data set with geocoded datasets about socio-economic and environmental health risk factors that might predict the best treatments for chronic disease with greater precision.

Max produces a number of findings that increase the precision with which physicians can select treatments, particularly for ethnic Samoans with asthma living in proximity to freeways and septuagenarians living at high altitudes with incident cancer.

Three years into the Exploratory Precision Medicine grant, Harvey Hacker at Computer Science University demonstrated that he could apply linear programming methods to uniquely identify two of the individuals in the Public Data set by combining it with voter registration records and the same geocoded data sets in use by Max Researcher. These two individuals represent 0.01% of the Large National Survey Population. Harvey alerts Large National Survey before he publishes his findings in Computer Science journal and as a New York Times Op Ed. Approximately 20% of the participants in the National Survey withdraw their data from the Large National Survey.

In response, the Large National Survey removes publically available online data set and changes agreements to require assurances from users that they will not combine either restricted or public use data with other data sets. Max Researcher destroys the data she has been using, shuts down her lab, and takes a job at Venture Capital Drug Discovery Firm, which uses privately brokered data sets with greater utility for precision medicine. Max cannot publish her findings under the auspices of her new organization until patents for precision treatments are granted.

**Questions:**

- What is the best way to manage funding agencies' mandates for data sharing with privacy concerns?
- What ethical obligations does Harvey have share or protect the algorithms he used? What obligations do cryptographers have to accurately communicate privacy risks? What obligations do scientific journals carry to publish or protect methods that might be used for unethical purposes?
- Should Large National Survey research participants be alerted of cryptographers' findings as newly identified risks?
- With the newly published algorithms, several other publically available research data sets are vulnerable: Should these data sets also be removed? Should participants be warned?
- The Large National Survey Data Stewards create a computing enclave where the Public Use data can be accessed and analyzed but not downloaded. The capacity limitations render many types of analysis infeasible. What alternatives exist?
- What standards or guidelines exist now for assessing tradeoffs between privacy risks and utility?
- Should researchers working under IRBs receive any special status or trust that would distinguish them from members of the public so that they might combine data sets to add value to the data?
- How can the risks of reidentification be balanced against the potential loss of valuable health insights that result from the removal of data sets from the public domain?
- How can the risks of reidentification be balanced against the burden of limited data access for researchers and potential loss of health insights (e. g., when a researcher removes themselves and their entire research program from the public domain)?

| Title | Response |
|---|---|
| Description | Under terms of funding from NIH, data sets collected with public funds must be made available while protecting privacy. Privacy researchers have shown that such data sets do not truly protect privacy, an issue that has received substantial public attention. This has resulted in more conservative approaches by data stewards, increasing barriers to data use by researchers. |
| Primary actor /participant | Researcher, Data Stewards |
| Support actor /participant | Funding agency |
| Preconditions | • Data Sharing Policies from funding agencies exist<br>• Participants have been consented<br>• Public data repository can be accessed and combined with other public data |
| Post conditions | • The researcher collects and analyzes the data for a specific research study.<br>• Data sets are removed from public access |

| | |
|---|---|
| Alternatives | • Cryptographers do not alert data stewards before results are released<br>• Cryptographers post code online and make it available for unrestricted use<br>• In addition to Drug Discovery, Venture Capital Drug Discovery Firm is selling data to marketers about the likely identities of individuals and their treating physicians for drug detailing. |
| Considerations | • Conflicts between mandates of funding agencies for data sharing and privacy concerns<br>• Conflicting interests between cryptography research publishing incentives, privacy of research participants, and<br>• Public perception of risk vs. actual risk. |
| Data Elements Considered | Survey, Laboratory, Demographic, and Geocoded Data about Environmental Risks, consented data from administrative claims |
| Purpose of the Data Collection | Precision Medicine |
| Purpose of Data Use | Research |
| Terms of Transfer to the Data Holders | Consent |
| Terms of Transfer to Researchers | IRB approval, Agreements negotiated with Data Stewards |