# Privacy-by-Design: Understanding Data Access Models for Secondary Data

**Hye-Chung Kum, PhD[1], Stanley Ahalt, PhD[1]**
**[1]NC-TraCS & Dept of Computer Science, University of North Carolina, Chapel Hill, NC**

**Abstract**

*Today there is a constant flow of data into, out of, and between ever-larger and ever-more complex databases about people. Together, these digital traces collectively capture our **social genome**, the footprints of our society. The burgeoning field of **population informatics** is the systematic study of populations via secondary analysis of such massive data collections (termed "big data") about people. In particular, **health informatics** analyzes electronic health records to improve health outcomes for a population. Privacy protection in such secondary data analysis research is complex and requires a holistic approach which combines technology, statistics, policy and a shift in culture of information accountability through transparency rather than secrecy. We review state of the art in privacy protection technology and policy frameworks from widely different fields, and synthesize the findings to present a comprehensive system of privacy protection in population informatics research using the privacy-by-design approach. Based on common activities in the workflow, we describe the pros and cons of four different data access models – restricted access, controlled access, monitored access, and open access – that minimize risk and maximize usability of data. We then evaluate the system by analyzing the risk and usability of data through a realistic example. We conclude that deployed together the four data access models can provide a comprehensive system for privacy protection, balancing the risk and usability of secondary data in population informatics research.*

**Introduction**

Today, nearly all of our activities from birth until death leave digital traces in large databases. Together, these digital traces collectively capture our *social genome*, the footprints of our society. Like the human genome, the social genome data has much buried in the massive almost chaotic data. If properly analyzed and interpreted, this social genome could offer crucial insights into many of the most challenging problems facing our society (i.e. affordable and accessible quality healthcare). The burgeoning field of *population informatics* is the systematic study of populations via secondary analysis of massive data collections (termed "big data") about people. In particular, *health informatics* analyzes electronic health records to improve health outcomes for a population. However, traditional approaches for privacy protection via informed consent and de-identification are no longer effective in an era where huge amounts of information can be synthesized from public data. On the one hand, true informed consent is impossible in secondary data analysis because the research question is not known at the time of data collection. On the other hand, de-identification has proven to be ineffective due to linkage attacks that can re-identify certain people. Clearly, building easily usable data that have low risk of disclosure is difficult. We need a better model for privacy protection in secondary data analysis that goes beyond anonymity and takes a more holistic approach[1]. This paper describes the cyber infrastructure required to conduct population informatics research in a coordinated, responsible manner that protects the privacy of individuals while allowing for meaningful insights.

Here, we propose a new paradigm that regards microdata about people as valuable but hazardous research material. Integrated microdata about people can hold the key to transforming biomedical sciences to new levels of evidence and investigation. Yet, when handled improperly, there is the potential for serious privacy violations that can undermine the public trust in research. Under this new paradigm, we take the *privacy-by-design* approach to privacy protection and focus on building a safe environment, consisting of secure computer systems and policy frameworks, in which data can be analyzed safely. Privacy-by-design goes beyond the narrow view of privacy as anonymity and attempts to meaningfully design privacy principles and data protection into the full system[2]. In this paper, we design a secure laboratory for secondary data analysis that incorporates the following two principles of privacy and usability. First, the *Minimum Necessary Standard* states that maximum privacy protection is provided when the minimum information needed for the task is accessed at any given time. Second, the *Maximum Usability Principle* states that data are most usable when access to the data is least restrictive (i.e. direct remote access is most usable).

A secure laboratory for secondary data analysis where safe research can take place with accepted protocols has three components. The laboratory is (1) a well-designed secure computer system, with (2) required secure software and data to carry out the research in a privacy preserving manner, along with (3) a policy framework for human subject protection in secondary data analysis of microdata about people. In this paper, we focus on the computer system, software, and data in terms of data access models with some discussion of the policy framework. In designing privacy into a comprehensive system of data access models, we reviewed state of the art privacy protection technology and policy frameworks from widely different fields–WWW[3], software management[4], social computing[5], statistics[6], and law[7] – and synthesized the findings. We design the system around the workflow based on the four

most common activities in secondary data analysis. The four corresponding levels of data access - restricted access, controlled access, monitored access, and open access – can offer optimum privacy protection while still providing maximum usability for the given data and activity. Together the four access levels can provide a comprehensive system for privacy protection for most secondary data analysis in population informatics research.

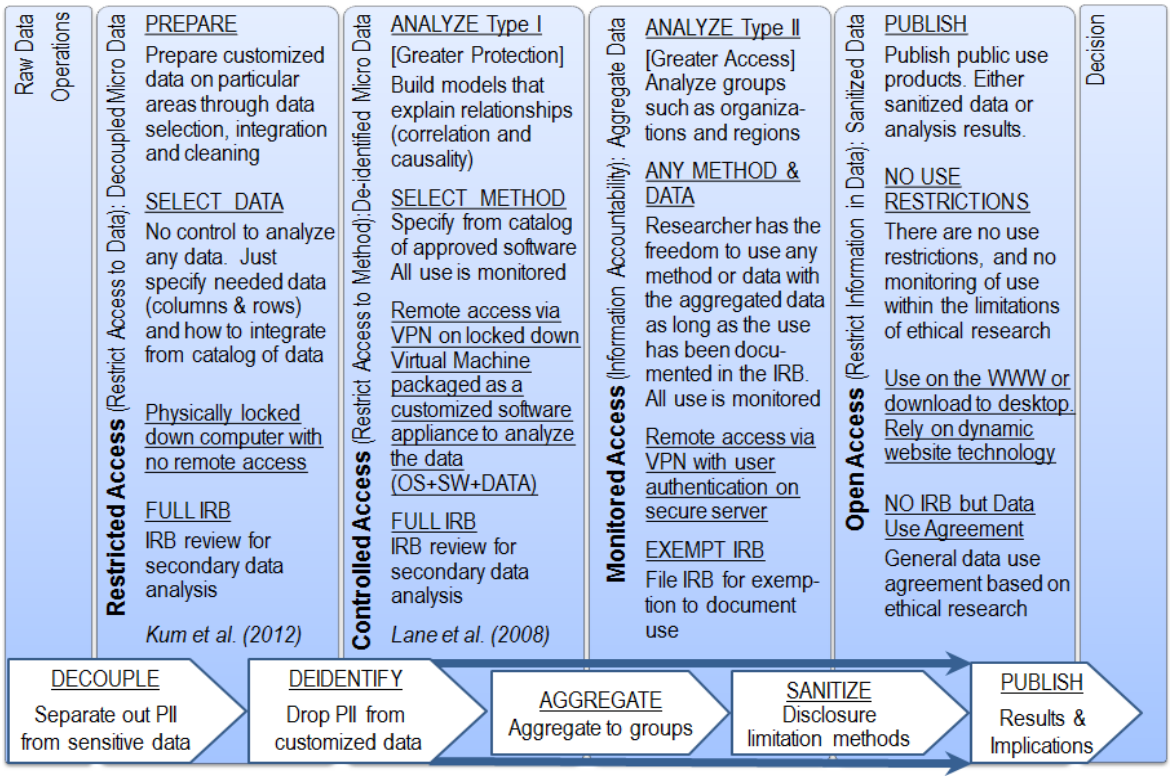


**Figure 1.** Data access models for privacy protection in population informatics research workflow (data to decision)

**Table 1.** Comparison of risk and usability (Usability 1.1 & 1.2 for Risk 1, Usability 2 for Risk 2)

| | | Restricted Access | Controlled Access | Monitored Access | Open Access |
|---|---|---|---|---|---|
| Example Systems | | US Census RDC | Secure Medical Workspace | Secure Unix Servers | Public Website |
| Type of Data | | De-coupled micro data including PII | De-identified micro data | Aggregate data | Sanitized data |
| Privacy Protection Methods Used | | • Encryption for decoupling<br>• locked down computer with physical restriction | Locked down VM to restrict software on the computer and data channels | • Information accountability<br>• Exempt IRB | Disclosure limitation methods |
| Monitor Use | | On and off the computer | On the computer | On the computer | No monitoring |
| Usability | U1.1: Software (SW) | Only preinstalled data integration & tabulation SW. No query capacity | Requested and approved statistical software only | Any software | Any software |
| | U1.2: Other data | No outside data allowed | Only preapproved outside data allowed | Any data | Any data |
| | U2: Access | No Remote Access | Remote Access | Remote Access | Remote Access |
| Risk | R1:Crypto-graphic Attack | Highly Difficult | Fairly Difficult. Would have to break into VM. | Easy to run sophisticated SW with outside data | NA |
| | R2: Data Leakage | Very difficult. Memorize data and take out | Physical data leakage (Take a picture of monitor) | Electronically take data off the system. | NA |

**System Design: Data to Decisions**

Broadly speaking, the purpose of secondary data analysis is to transform raw data into insights to inform decision making. Raw data can come from hospital patient records, government programs such as Medicaid, or many other sources. Researchers then interpret the data by applying analytical methods to gain insights. These insights are then shared to inform decisions about, for example, medical treatments, hospital policies, insurance policies, government programs, or legislation. The route from data to decisions includes several important steps. In the proposed paradigm, these steps are broken into four workflow elements (data preparation, analysis of microdata, analysis of aggregate data, and publication) with four corresponding levels of data access (restricted, controlled, monitored, and open). Not every research endeavor requires all four steps or levels of access; the type of data needed and the appropriate level of access depends on the intended use of the data. However, taken as a whole, we believe the proposed system

provides a more robust infrastructure for both privacy protection and data access than the approaches that are currently most common in biomedical research, which is largely dominated by approaches that either sacrifice privacy for greater data access (for example, "trusted" researchers are given full access to de-identified microdata) or protect privacy at the expense of data quality (for example, many integrated data do not properly account for record linkage errors). The elements of this system are described below in reverse order from sanitized data because most people are familiar with open access and monitored access (figure 1 and table 1).

**1. Open access to sanitized data: Statistical disclosure limitation methods to restrict information in data**

Direct publication of data to the WWW can greatly increase its usability because any researcher may access the data at any time for any purpose. In an *open access* system, the data are typically sanitized before release in order to protect privacy. The statistical disclosure limitation research investigates safe methods to release unrestricted access to the data by restricting the information disclosed in the data rather than access to the data. These methods to sanitize data are the most researched approach to privacy protection[6]. The main limitation of open access data publication is that sanitized data are difficult to use for data integration and repurposing. The process of data sanitization requires making assumptions about the data, potentially causing unexpected side effects and misinterpretation. For example, top coding public use data can result in incorrect information being provided to policy makers, which can ultimately lead to misinformed decisions and social harm[8]. As a result, we recommend that even sanitized data be published with data use agreements and a general code of conduct including proper use of the data from both a privacy perspective and a data methodology perspective.

**2. Monitored access to aggregated data: Information accountability through user authentication**

In *monitored access*, data are stored on a secure server and authorized researchers access the data by logging into the server over a secure VPN connection. It is the most common form of data access in research today. The main mechanisms for privacy protection are data encryption, secure VPN connection, and user authentication. User authentication technology has developed from simple password protection to using dynamically generated RSA keys. Monitored access implements the information accountability[3] approach by making access easy for authorized users, but keeping logs of user activity on the server, which can be investigated if a data breach is suspected. Under HIPAA, if identifiable health records are breached, it can lead to serious consequences. Thus, we recommend that monitored access is only used for raw aggregate data about groups of people and not individual level microdata. In sum, information accountability is a shift in the culture of digital privacy from using technology to support secrecy (hiding information) to using technology to support transparency (keep logs of activity and make it difficult to alter logs). Such an approach to digital privacy aligns well with the legal premise of privacy as contextual integrity which dictates that privacy is contextual and depends on agreed norms of expectation for privacy[7]. When there are agreed norms for privacy protection, and reliable technology can hold parties accountable through transparency, digital privacy becomes easier to enforce in the open environment. In an academic setting where reputation and peer review is the norm, accountability through transparency is the best method to enforce ethical behavior.

Compared to controlled access, monitored access gives users much greater flexibility when using data. Researchers are free to install or write their own analysis software, or to bring their own data into the system for linkage to investigate other factors. The cost of this increase in data usability is a higher risk to privacy due to an open server in which authorized users have full control of the machine. There is little protection against negligent or malicious insider attack. The only mechanism for protection is via transparency based on the exempt IRB. It will be important to file the IRB because the process will explicitly self-define the scope of data use. Then the log of all user activity and the IRB will provide the full transparency required to enforce information accountability when a breach is suspected. In addition, there is exposure to potential malware on the PC that is being used for remote access. PCs used for remote access typically have a high risk to malware because they are used to browse the web. When these questionable systems are used to remotely access the server, the monitored access server and the sensitive data are potentially exposed to vicious unintentional threats. In comparison to open access, researchers can freely repurpose the aggregate data for population informatics research without worrying about inadvertent errors in the results.

**3. Controlled access to de-identified microdata: Specialized software appliance using virtual machines**

Once a customized dataset is produced, researchers can analyze the data at the level of individuals (microdata) or groups (aggregate data). The research task that scientists spend the most time on is analyzing customized data. Thus, direct remote access is crucial for this step. The *minimum necessary* amount of data needed for analyzing microdata is de-identified individual data. De-identified data is not fully protective of privacy because linkage attacks using quasi-identifiers can be used to re-identify individuals. Therefore, we recommend a *controlled access* system for this step. Controlled access is a remote access system that restricts all activities on the computer using dynamic, configurable, role-based access policies that are enforced automatically and fully monitored with an audit system. Controlled access is a form of access that balances the pros and cons of restricted access (privacy protection but high

barriers to access) and open access (easy access but little privacy protection). Both monitored access and controlled access try to balance between the two extremes, but monitored access is biased toward easier access while controlled access is biased toward privacy protection.

A controlled access system can be managed through remote access via VPN connection combined with virtualization. Using this setup, the customized dataset and pre-approved, customized data analysis software are shipped as a virtual machine (VM)[4] to the researcher's desktop. Scientists can only use data analysis tools that have been preselected from a library and provided for them in the custom built appliance. Because all data channels are locked down, no data can be brought into or taken off the appliance, making it difficult to perform linkage attacks and significantly reducing unintentional data loss. In addition, even when the PC used for remote access is compromised, malware cannot access any part of the VM because the VM is isolated from the host OS even though it shares the same hardware. At worst, the VM will not launch due to problems on the host OS with no compromise to confidential data. This essentially eliminates the two risks discussed in the monitored access system section. There are a few working prototypes of controlled access systems[9]. The main privacy threat in a controlled access system is physical data leakage (for example, by taking a picture of the screen). Thus, it is not secure enough to use with personally identifiable information (PII), but is sufficient for use with de-identified microdata.

Along with these technical protections, we recommend full IRB processes for the use of de-identified microdata because it is impossible to fully anonymize large microdata for all subjects. The full IRB processes should balance the benefits of research with the potential for harm to the human subjects from secondary analysis. The best ways to assess potential harm would be to evaluate the risk of confidential attribute disclosure given the table of attributes used for the study and the computer system that will be used to analyze the data. It will be important to train more researchers to become skilled in spotting potential harm and differentiating between required data and extraneous data for typical biomedical research over time. Once the analysis is completed, following standards for publication of research results would be sufficient protection against data released from the controlled access system. Besides statistical analysis, there are two other common activities that occur on microdata. Scientist can prepare aggregate datasets to be analyzed in monitored access systems or sanitize the microdata to build public use data, which can be released to an open access system.

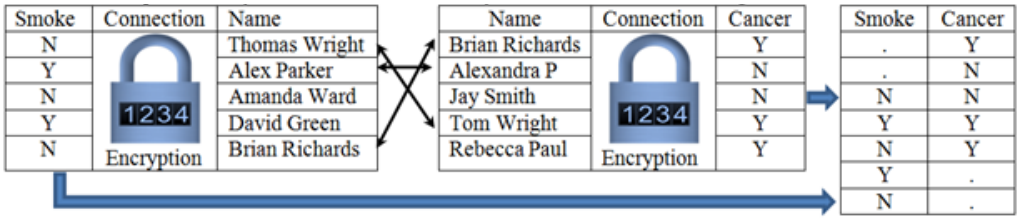


**Figure 2.** Decoupled and chaffed data allows for data integration without sensitive attribute disclosure

## 4. Restricted access to decoupled microdata: physically restrict access to data

To use raw data for secondary data analysis, researchers must first prepare a customized dataset for a particular research question. Data preparation can entail integrating data from different data sources, selecting attributes specific to the intended purpose, cleaning the data to ensure consistency and eliminate duplication, and selecting a sample. These activities require wide access to lots of raw data that often includes personal identifying information (PII), presenting a high risk to privacy. Currently, access to PII for data preparation is restricted through collaborative agreements in which trusted parties inside organizations housing raw data (such as staff at hospitals or state health statistics departments) prepare the data and provide researchers de-identified data. However, relying on such staff members to prepare customized datasets can make it difficult for researchers to control error and ensure data are properly integrated and cleaned.

In the proposed model, researchers may be given direct access to raw data via established encryption technology that decouples PII, such as name, from sensitive data, such as health status (see Figure 2). For effective data integration, researchers typically require PII, but not the information connecting PII to sensitive data; thus, decoupled data are the *minimum necessary* amount of information required for this step. Although PII remains in the dataset, this carries a low potential for harm because identity is not linked to attributes containing sensitive information. Although attribute disclosure occurs mostly through identity disclosure, it is important to distinguish between the two because identity disclosure *without* attribute disclosure has low potential for harm[5, 6]. Kum at el present the details of a decoupled data system that provides privacy-preserving data integration with error management[5]. They show that in a decoupled data system, attribute disclosure is fundamentally blocked using encryption and chaffing, adding fake data. Even identity disclosure is rare when chaffing, shuffling, and isolation of fields are properly used. Using a decoupled system can significantly reduce data loss by authorized users resulting from both unintentional and deliberate behavior which represent a significant threat to privacy.

A key characteristic of the restricted access decoupled system is that the scientists have very limited control to manipulate the data. The scientists interact with the computer much like they might with an internal collaborator by specifying the data they want to prepare using the metadata for the hospital records. Then the bulk of the work is done by the decoupled data integration software. It is only when the software runs into ambiguous decision points that the scientist is required to provide guidance on the decision based on the information the software provides (i.e. the difference of two PIIs). Scientists can run frequencies and cross tabulations that are not PII to assist in attribute and sample selection, but they cannot query or view the identifiable microdata independently. Given that decoupled data allows direct access to PII, we recommend *restricted access* for data preparation. Restricted access is a highly secure system with full monitoring of all activities, on or off the computer at the cost of high barriers to use the data. It does not allow remote access, so scientists must travel to designated locations to use locked-down computers. All releases of the data from the system are restricted, including printouts. One such setting is the RDC (research data centers) at the U.S. Census. Researchers should only need to work in such restrictive settings for brief periods.

**Data Risk and Usability Evaluation through Example**

We evaluate the proposed system using a real example in which Yung at el. linked the New York central cancer registry with Medicaid enrollment and claims files to assess cancer care among the poor[10]. Table 2 shows the data usability and privacy risk of conducting data preparation and microdata analysis for this research in two different environments. The results show that restricted and controlled access systems, which are locked down computer systems, can allow secondary analysis on microdata to be carried out with better usability of data as well as reduced risk of harm to the subjects of the data.

**Table 2.** Analysis of risk and usability of the conventional system and the proposed system

| Workflow | Data Preparation (Data Integration and Selection) | | Analysis of Micro Person Level Data | |
|---|---|---|---|---|
| System | Conventional System | Proposed System | Conventional System | Proposed System |
| Model | Indirect Access via Health Dept. | Direct Restricted Access | Monitored Access | Controlled Access |
| Access | No direct access to data | Direct access to data | Remote direct access for authorized users | |
| Type of Data | Multiple Identifiable Microdata Tables | Multiple Decoupled Microdata Tables | De-identified integrated microdata | De-identified integrated microdata with P(linkage) |
| Analysis of Risk and Usability | The proposed model **increase data usability** by allowing researcher to directly carry out record linkage and data (attribute and sample) selection leading to more accurate analysis by propagating the error, probability of linkage, to the analysis phase. **Reduce insider attack risk** by decoupling PII from the sensitive data through encryption. | | The proposed model r**educe risk significantly from insider attack and malware** in the proposed model by (1) restricting activities on the VM, and (2) running the VM in isolation from the host OS | |

**Conclusion and Future Work**

There is a direct relationship between data usability and risk to privacy; greater access to data generally leads to a higher privacy risk and more restricted access generally provides better privacy protection[11].Thus, protecting privacy in secondary data analysis requires carefully balancing these factors through a holistic approach involving technology, policy, statistics, and a shift toward a culture of information accountability rather than secrecy. In this paper, we synthesized data access models from a variety of fields to propose a system of four data access models that, taken together, can improve both data access and privacy protection for use of secondary data in population informatics research. Future work may build upon this new paradigm by investigating ways to more effectively build transparency into the use of secondary data, such as by making approved IRB submissions public, notifying the public of research data and offering a chance to opt out, or developing and training of a code of ethics around understanding the obligations to the confidential relationship between the researcher and the subjects of the secondary data.

**References**


1. Lane J, Heus P, & Mulcahy T. Data access in a cyber-world: Making use of cyberinfrastructure.. Transaction on data privacy. 2008;2-16.
2. Shapiro S. Inside risks - Privacy by design: Moving from art to practice. Comm. of the ACM 2010;**53**:6.
3. Weitzner DJ, Abelson H, Berners-Lee T, et al. Information accountability. Comm. of the ACM 2008;51:82-7.
4. Sapuntzakis C, Brumley D, et al. Virtual Appliances for Deploying and Maintaining Software. LISA USENIX 2003;181-94.
5. Kum, H.C., Ahalt, S, Pathak, D. Privacy Preserving Data Integration Using Decoupled Data. SPSN Springer 2012.
6. Fienberg SE. Confidentiality, privacy and disclosure limitation, Encyclopedia of Social Measurement, AP 2005;**1**:463-9.
7. Nissenbaum HF. Privacy as Contextual Integrity. Washington Law Review 2004;**79(1)**:19-158.
8. Lane J. Optimizing the Use of Micro-data: An Overview of the Issues Presented at I Quality and Access to Federal Data. ASA Section on Government Statistics 2005.
9. Shoffner, M, & Mostafa, J. Secure Medical Research Workspace. CTSA Informatics KFC 2012 Annual Meeting; Chicago.
10. Yung RL, Chen K, Abel GA, et al. Cancer disparities in the context of Medicaid insurance: a comparison of survival for acute myeloid leukemia and Hodgkin's lymphoma by Medicaid enrollment. Oncologist 2011;**16(8)**:1082-91.
11. Duncan G, Keller-McNulty S, Stokes S. Disclosure risk vs. data utility: The R--U confidentiality map. NISS TR-121 2001