Privacy by Design for Secondary Data Analysis

Hye-chung Kum, kum@email.unc.edu
Department of Computer Science, School of Social Work, UNC-CH

Stanley Ahalt, ahalt@renci.org
RENCI, Department of Computer Science, UNC-CH

***Abstract*—Objective** Secondary data analysis is becoming more powerful and commonly utilized for biomedical research using patient records and genomic data. In both data, de-identification has been proven to be ineffective due to linkage attacks that can re-identify some subpopulation of the data. We need a better model for privacy protection in secondary analysis of biomedical data. **Design** In this paper, we review state of the art privacy protection technology and policy frameworks from widely different fields – WWW, software management, social computing, statistics, and law – and synthesize the findings to present a comprehensive model of privacy protection in biomedical research using the privacy by design approach. Based on common activities in the research pipeline, we propose four different data access systems that minimize risk and optimize utility in data. We then evaluate the model by analyzing the risk and utility of data through a realistic example. **Results** We found that there are four common types of activity in the research pipeline that require different levels of data and protection – decoupled microdata, de-identified microdata, raw aggregate data, and sanitized data. The four corresponding levels of data access – restricted access, controlled access, monitored access, and open access – together can provide a comprehensive model for privacy protection, balancing the risk and utility of secondary data analysis for biomedical research. **Discussion and Conclusion** Privacy protection is a complex issue and requires a holistic approach combining technology, statistics, policy and a shift in culture of information accountability through transparency rather than secrecy.

Keywords- privacy by design, secondary data analysis, open access, monitored access, controlled access, restricted access

I. INTRODUCTION

Traditional approaches for privacy protection via informed consent and de-identification are no longer effective in an era where access to large amounts of public data is virtually effortless and computational methods exist to synthesize huge amounts of information from it. Secondary data analysis is becoming more powerful and commonly utilized for biomedical research using patient records and genomic data. On the one hand, 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 some subpopulation of the data in both medical records and genomic data [1, 2, 3]. Clearly, building data with utility 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 [4, 5].

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 a new level of evidence and investigation. Yet at the same time, 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 (figure 1). 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 from the beginning of the development process to deployment, use, and ultimate disposal [6]. In this paper, we design a secure laboratory for secondary data analysis, that incorporates the following two important principles of privacy and utility:

- Minimum Necessary Standard: Maximum privacy protection is provided when the minimum information needed for the task is accessed at any given time.
- Maximum Utility Principle: Maximum utility of data results from direct access to the required information when needed

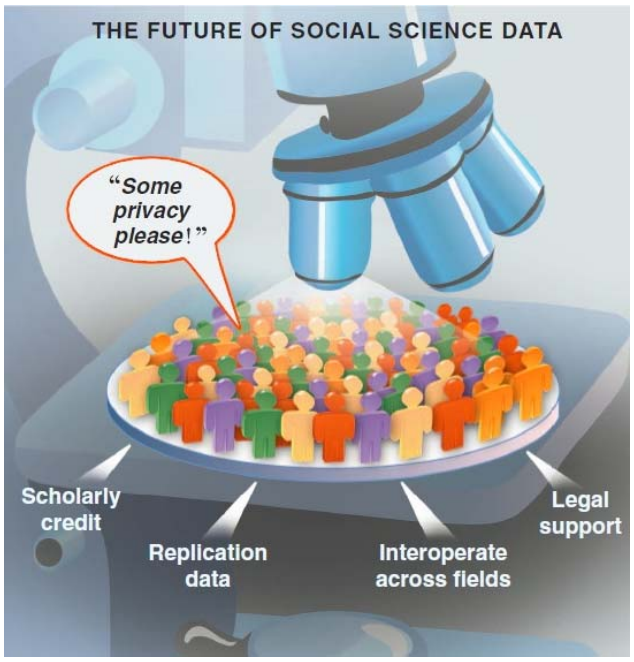

Figure 1. Microdata about people as valuable but hazardous research material that requires a safe lab [13]

A secure laboratory for secondary analysis of microdata where safe research can take place with accepted protocols has three basic 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 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. In our conclusion, we discuss directions for future work for the policy required to make such a system fully functional.

In designing privacy into the comprehensive data access models, we reviewed state of the art privacy protection technology and policy frameworks from widely different fields – WWW [7, 8], software management [9], social computing [10], statistics [11], and law [12] – and synthesized the findings. We build the system around the pipeline for research using secondary data based on the kinds of data required for certain research activities. We found that there are four types of data associated with common activities in secondary data analysis – decoupled microdata (preparing data), de-identified microdata (analyzing microdata at person level), raw aggregate data (analyzing aggregate data at group level), and sanitized data (publishing for public consumption). Each type of data has a different level of privacy threat and requires a different level of protection. Thus, we design four corresponding levels of data access - restricted access, controlled access, monitored access, and open access – which can offer optimum privacy protection while still providing maximum utility for the given data and activity. Together the four access levels can provide a comprehensive model for privacy protection for most secondary data analysis.

II. Data Access Models for Privacy Protection in Biomedical Research

We analyzed the biomedical research pipeline into the minimum information required to perform certain tasks following the minimum necessary standard . Then for each data type, we designed the least restrictive access, as per the maximum utility principle, to the data that can still maintain privacy. The

types of data used for biomedical research in secondary analysis depends on how much pre-processing has been done from the original source. For simplicity, we will focus the discussion on hospital records, but much of the discussion can be extended to other sources. Figure 2 depicts the flow of raw data from hospital records to sanitized data for public use, and Table 1 compares the risk and utility of data in each access model. We discuss the different systems backwards from sanitized data to decoupled data because most people are familiar with open access and monitored access systems.

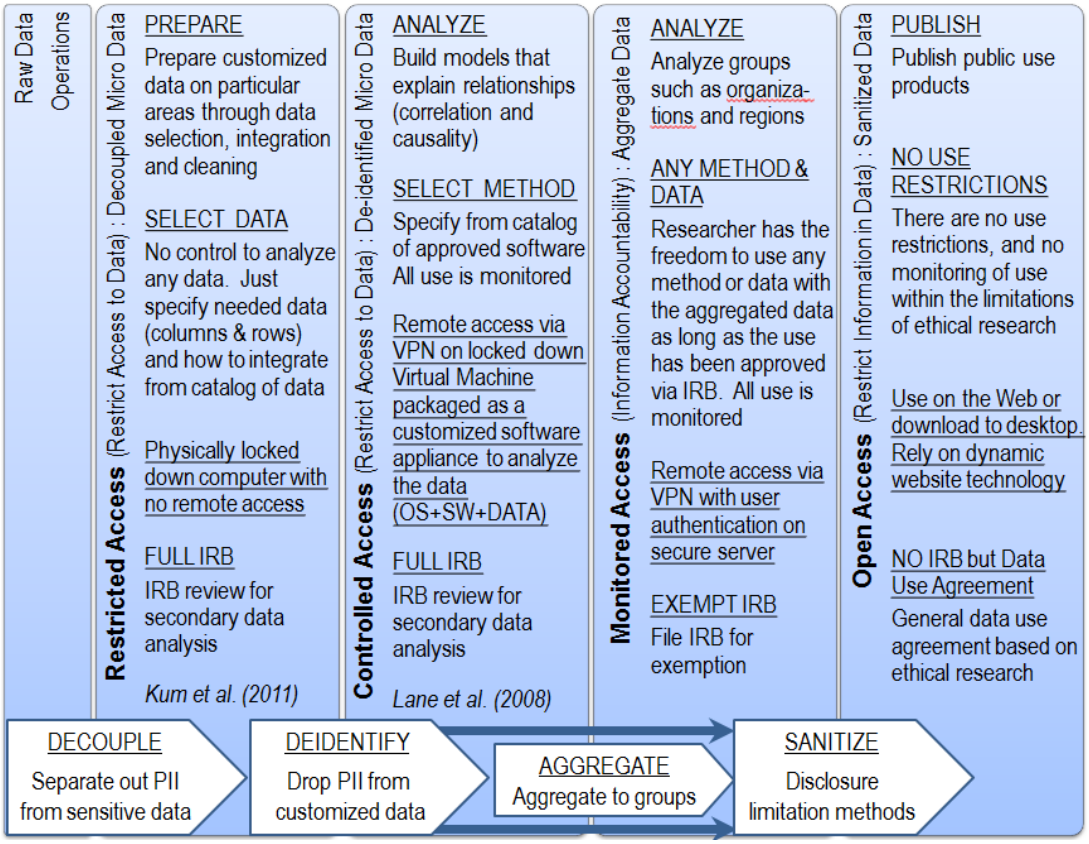

Figure 2. Comprehensive model of privacy protection in the research pipeline

		Restricted Access	Controlled Access	Monitored Access	Open Access
Example Systems		US Census RDC (Research Data Center), CDC RDC	Secure Medical Workspace NORC data enclave	Most conventional secure Unix servers	Census 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 • locked down computer with physical restriction	Locked down VM to restrict software on the computer and data channels	• Information accountability • Exempt IRB	Disclosure limitation methods
Monitor Use		On and off the computer	On the computer	On the computer	No monitoring
Utility	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: Cryptographic 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

Table 1. Comparison of Risk and Utility (Utility 1.1 & 1.2 for Risk 1, Utility 2 for Risk 2)

Z -> AZB+C	where A is a matrix that operates on the rows, B is a matrix that operates on the columns, and C is a matrix that adds perturbations or noise.

Figure 3. Data masking expressed as matrix math (Equation 1)

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

Open access is the easiest form of access. It is achieved by simply publishing data on the WWW, such as the many instances of public use data available on the web today. The downside is that there is little privacy protection, so these public use data are sanitized before release. 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. In short, these methods attempt to limit disclosure by changing the data where the chances of disclosure are estimated to be high. The static methods transform the data before releasing it, while more recent work in query systems keeps track of the queries made and transforms the data dynamically as needed [14, 15, 16, 17]. Most static methods can be described as either data masking or simulation. Data masking methods (i.e. generalization, suppression, adding noise) can generally be expressed using equation 1 (figure 3). Recent work in computer science on differential privacy and k-anonymity also fit under this model, but in addition, discuss the required properties of the released data that needs to be met for privacy guarantees [18, 19, 20]. Simulation methods first model the data then attempt to generate realistic synthetic data. It is important to note that even fully simulated data has potential privacy issues because some simulated individuals may be virtually identical to real people. A good overview of these methods can be found in [11, 15]. We note that even sanitized data should be published with data use agreements around a general code of conduct that users must agree to before downloading the data. Those should include proper use of the data methodologically as well as for privacy reasons, because incorrect use of data, a greater hazard for sanitized data, can lead to other harm.

The main limitations with these methods is that masked or synthetic data are difficult to use for data integration and repurposing. When data are masked or simulated there are hidden assumptions as to when results are accurate, making flexible repurposing difficult. Unexpected side effects, either detected or not, can occur when these unknown assumptions are not upheld. For example, top coding public use data can result in incorrect information being provided to policy makers which can ultimately lead to incorrect decisions [21].

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

We use the term *monitored access* to refer to the most common form of data access in biomedical research today. In monitored access, data is stored on a secure server and authorized researchers access the data by logging into the server over a secure VPN connection. 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 approach by making access easy for authorized users, but keeping logs of user activity on the server. When a data breach is suspected, security forensic specialists investigate. If the specialists cannot determine exactly what data was breached, institutions are required to assume all data was breached. Under HIPAA, for health records such situations can lead to huge financial penalties if the data is identifiable. Thus, we recommend that monitored access is only used for raw aggregate data about groups of people and not 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 on norms of expectation for privacy [12]. When there are agreed on norms for privacy, 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, there are two important ways that data access is easier, resulting in better data utility. First, for authorized users, the computer is an open system with little restriction on

what software researchers can use to analyze the data. Researchers are free to install or write their own software, or bring in their own data into the system for linkage. Second, only an exempt IRB has to be approved. It is still important to file the IRB because the process will explicitly self-define the scope of data use. In fact, the only mechanism preventing researchers from taking data off the system will be that it was not stated in the IRB. 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 comparison to open access, researchers can freely repurpose the aggregate data for biomedical research without worrying about inadvertent errors in the results.

The cost for this increase in data utility is that no person-level data can be accessed because there is higher risk to privacy due to an open server where authorized users have full control of the machine. The seemingly small risk is exacerbated by the 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.

Hence, monitored access is designed for aggregated data that were built using the controlled access system. Aggregated data represent data for a larger unit of analysis, such as organizations or regions, compared to microdata where the unit of analysis is a person. Risk of harm to a person in aggregated data is different from disclosure risk in microdata. In general, the risk of harm is much less than in microdata but there are still issues to consider. First, aggregate data that represent a group of one person can lead to disclosure of personal information by linking files. Similarly, there is a marginal risk of disclosure for aggregate data that represent a small group of people, if enough information about some subgroup is known [22]. Second, regardless of group size, attributes that hold true for all members of the group have a risk of disclosure via membership in the group [11]. For example, publishing the maximum wage for a certain group at an institution will disclose that all members of the group make less than the published amount. Third, sensitive information about organizations can harm particular individuals in the organization (i.e., financial information about a hospital can harm the director). There is also the issue of privacy of organizations. The issue of what is confidential data and public information for organizations is different from privacy protection for individuals. We do not address this issue here, but acknowledge the need for a balance of privacy for organizations with the need for better transparency for accountability of organizations. These risks in data require that even aggregate data be analyzed with care in a monitored access system. As with microdata, we can also sanitize aggregate data for release to open access systems.

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

The research task that scientists spend the most time on is analyzing customized data. Thus, direct remote access is crucial for this step. Following the minimum necessary standard, scientists can work solely with de-identified microdata. External files are generally considered the largest threat to disclosure in de-identified data [22]. Thus, we recommend a *controlled access* system for analyzing de-identified microdata, which essentially restricts all activities on the computer. Controlled access is a remote access system that has dynamic configurable role-based access policies that are enforced automatically and fully monitored with an audit system. Remote access via VPN connection combined with virtualization can now provide the level of flexibility and security required to manage such a dynamic 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 extreme, but monitored access is biased toward easier access while controlled access is biased toward privacy protection. The secure workspace funded by NIH [23], the data enclave at NORC [5, 24], and the virtual center project funded by DHS [25] have built working prototypes for this type of computer system.

Basically, given some input, a customized software appliance is built per user with a locked down OS, requested and approved software for the analysis, and customized data prepared at a restricted access level, then shipped as a virtual machine (VM) to the scientists' desktop. That means the scientists can

only use the data analysis tools, preselected from a library and provided for them in the custom built appliance. Furthermore, with all data channels locked down, no data can be brought into or taken off the appliance reducing unintentional data loss significantly. Furthermore, the most common form of re-identification threat via linkage attacks with external files becomes very difficult since no new data could be brought into the system. In addition, even when the PC used for remote access is compromised, the malware cannot access any part of the VM, because the VM is totally 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 in the monitored access system. The main threat of controlled access is the possible physical data leakage (i.e. taking a picture of the screen). Thus, it is not secure enough for personally identifiable information (PII), making data integration at this level of access impossible.

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 a monitored access system for analysis of groups of people. Or scientists can sanitize the microdata to build public use data, which can be released to an open access system.

D. Restricted access to decoupled microdata: physically restrict access to data

The first activity in the research pipeline is to prepare a customized dataset for a particular research question. This typically requires integrating one or more data, selecting the attributes needed, and then selecting the sample to investigate. These activities require wide access to lots of data including PII which have a high risk of disclosure. Most research involving secondary data will need to access PII in some way to integrate data and prepare the customized data for analysis. Currently, access to PII for data integration is typically gained indirectly through collaborative agreements with trusted inside parties such as hospital staff or state health statistics departments where they will prepare the customized data then share the de-identified data. This leads to difficulty in controlling error and data cleaning during the data preparation step. In our model, we increase utility of the data by giving direct access to the data via established encryption technology to decouple the data into PII and sensitive data. Decoupling takes away the connection information from the PII to the sensitive data, which is not required for data integration following the minimum necessary standard. Figure 4 depicts three different levels of information that scientists can access. For good data integration, the researcher requires PII, but not the connection information from PII to sensitive data. Thus, decoupled data is the minimum amount of information required for this step. 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 [11]. Kum at el present the details of a decoupled data system that can provide privacy preserving data integration with error management [10]. They show that in a decoupled data system, attribute disclosure is fundamentally blocked using encryption and 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.

Given direct access to PII in the decoupled data, we recommend *restricted access*, the most constrained form of data access, for data preperation. 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. With no

remote access, scientists must physically go to designated locations to access locked down computers, and all release of the data from the system including print outs are restricted. One example of such a setting is the RDC (research data centers) that the US Census Bureau maintains for access to confidential microdata for research [26, 27].

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). The scientists cannot query or view the identifiable microdata independently. However, they can run frequencies and cross tabulations on the decoupled data that are not PII to assist in attribute and sample selection.

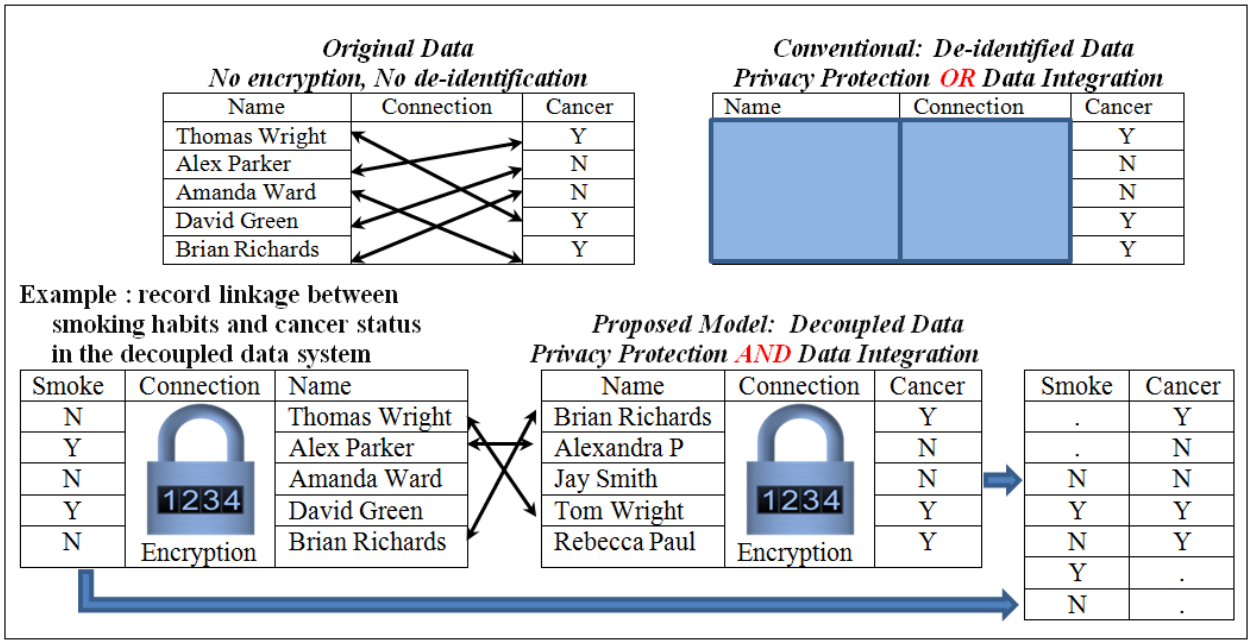

Figure 4. Three modes of information sharing in data

III. DATA RISK AND UTILITITY EVALUATION THROUGH EXAMPLE

We will evaluate the model for data risk and utility using a real example in cancer research using the cancer registry (CR), Medicaid data, and death records. Yung at el. linked the NY central cancer registry with Medicaid enrollment and claims files to assess cancer care in the economically disadvantaged population [28, 29]. We develop a realistic hypothetical example combining our own experience with their papers to evaluate the model. We use the example to analyze the risk and utility of data under two environments, one in which research is carried out in a conventional setting and the other in which the research is carried out using the proposed model in a hypothetical setting. Conventional methods of carrying out health informatics research is defined as monitored access on a secure server in collaboration with health departments. In comparison, we assess the change in risk and utility of data when we use decoupled data on restricted access servers and de-identified microdata on controlled access servers. Note that the discussed potential risk does not reflect the risk in any particular research setting.

The Yung at el. study links CR data with Medicaid and death records in two states, NY and CA, using probabilistic record linkage on SSN, name, DOB, and gender. It then uses Medicaid enrollment as a proxy for social economic status (SES) and studies the survival rate and hazard ratios for sociodemographics and clinical factors for two kinds of cancer, acute myeloid leukemia (AML) and

Hodgkin's lymphoma (HL) in the nonelderly population. They found that poverty does not seem to affect survival rate of AML patients, but it does affect the survival rate of HL patients. They suggest that the disparity in survival rate in HL patients is due to the complex outpatient treatment needed in HL which is not needed in AML. Here, we only focus on the NY analysis of the HL patients for brevity.

Figure 5 is the scenario for doing the full analysis in the two different settings. As seen in the figure, the scientist has direct access (HIGHER UTILITY) to much more data in the proposed model leading to three important ways in which the utility of the data is increased (1) for doing record linkage (UTILITY 1), (2) selecting attributes and samples (UTILITY 2), and (3) carrying out more accurate survival analysis (UTILITY 3). At the same time, risk is reduced. First, the three risks in the conventional setting are all eliminated in the proposed model by restricting activities on the VM (RISK 1, REDUCED RISK 1), running the VM in isolation from the host OS (RISK 2, REDUCED RISK 2), and decoupling PII from the sensitive data through encryption (RISK 3, REDUCED RISK 3). Second, recognizing the high risk in PII data, all activities on and off the computer are monitored and restricted including no print outs leaving the facility (REDUCED RISK 4). A more detailed description of the comparison is given in the appendix. We also extend the model to describe the difference in how monitored and open access systems are used in our model. In sum, we believe that more data can be safely analyzed in both the restricted access and controlled access systems, which are locked down computer systems, compared to the conventional monitored access system, which is an open computer system. Thus, by utilizing the restricted access and controlled access systems, most secondary analysis on microdata can be carried out with better utility of data as well as reduced risk of harm to the subjects of the data in the proposed model. In comparison, monitored access systems are well-suited for doing analysis on aggregated data which have lower risk of harm to individuals.

IV. Conculsion and Future Work

There is a direct relationship between the risk and utility of data for research. When risk is reduced in data, less information is shared with the researcher, and the utility of data is reduced [30, 31, 32, 33]. Thus, privacy protection in biomedical research using secondary data requires carefully balancing the risk and utility of data through a holistic approach that utilizes technology, policy, statistics, and a shift in culture of information accountability rather than secrecy. In this paper, we took the privacy by design approach to design a comprehensive model for privacy protection in secondary data analysis based on four common activities that occur in the research pipeline. The four data access models – restricted access, controlled access, monitored access, and open access – corresponding to each activity that was reviewed for risk and utility of data to evaluate the proposed model. We found that compared to the conventional setting for carrying out biomedical research, the new model provides both higher protection from re-identification and insider attacks and better access to data for higher utility of data. More research is needed in building transparency into research using this model such as publishing all approved IRB online, notifying the public of the existence of research data with easy mechanisms for opting out, as well as 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.

Acknowledgment

The model proposed in this paper combines many conversations authors had over the years. We cannot list them all, but would like to thank those that gave direct feedback to the paper as it was taking shape. We thank Jisung Kim, Gautam Sanka, Darshana Pathak, Rebecca Wells, Elliott Smith, Michele Weigle, Mike Reiter, Fabian Monrose, and Ren Bauer for their insightful comments. This research was supported in part by funding from the NC-DHHS.

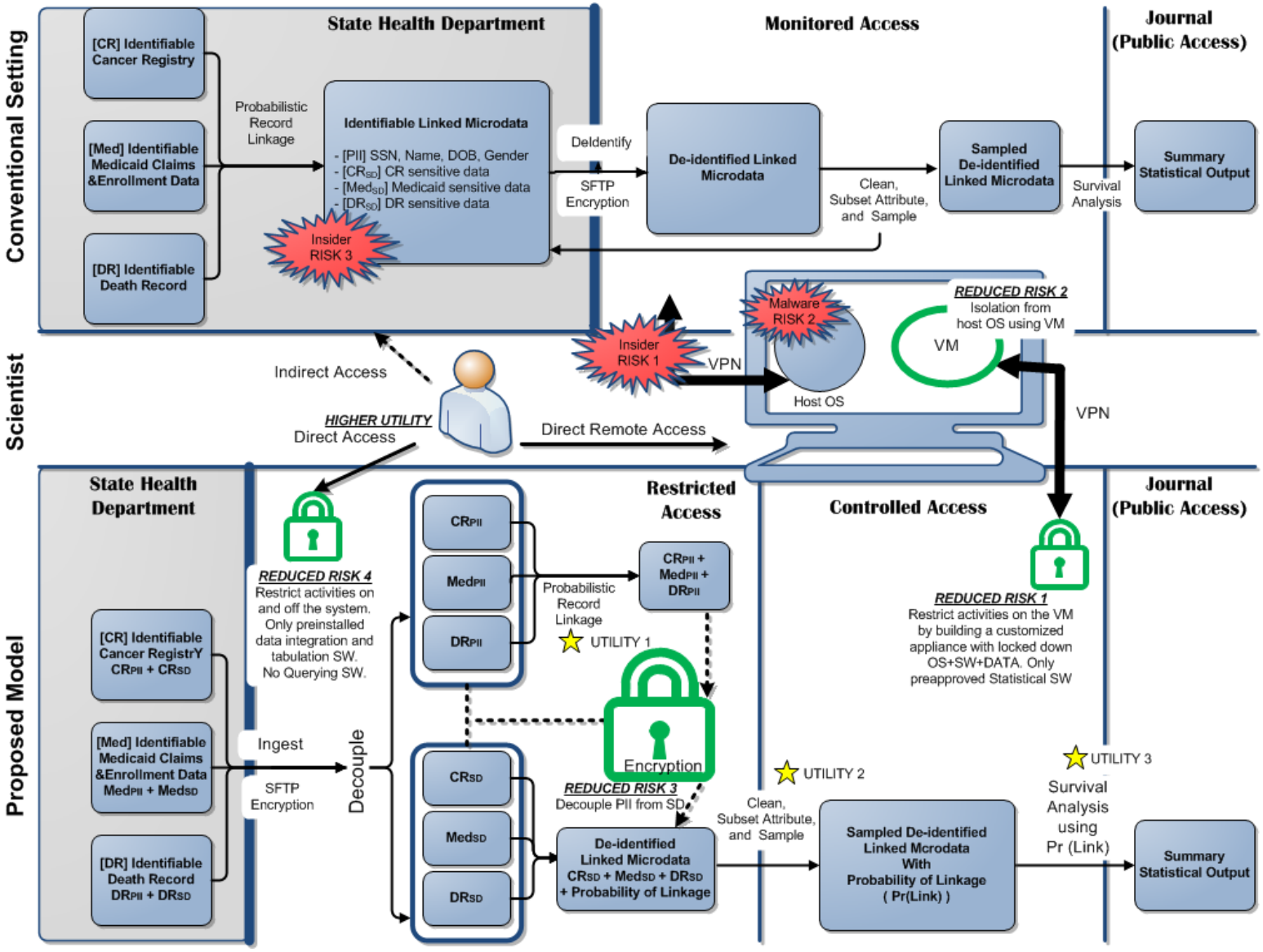

Figure 5. Risk and Utility Analysis: The proposed model results in *HIGHER UTILITY* due to direct access which allows for three benefits (UTILITY 1-3) and eliminates all three RISKs in the conventional model by restricting activities on the VM (*REDUCED RISK 1*), running in isolation from the host OS (*REDUCED RISK 2*), using encryption to decouple the PII from the sensitive data (*REDUCED RISK 3)* and by restricting activities on and off the restricted access system.

REFERENCES

1. Narayanan A, Shmatikov V. Robust De-anonymization of Large Sparse Datasets (How to Break Anonymity of the Netflix Prize Dataset). S&P (Oakland) 2008.
2. Sweeney L, Weaving technology and policy together to maintain confidentiality. Journal of Law, Medicine and Ethics 1997;**25**:98-110.
3. P^3G Consortium, Church G, Heeney C, et al. Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection PLos Genet 2009;**5(10)**:e1000665:doi:10.1371/journal.pgen.1000665
4. Narayanan A, Shmatikov V. Myths and fallacies of personally identifiable information. Comm.of the ACM 2010;**53**:24-6.
5. Lane J, Heus P, Mulcahy T. Data access in a cyber-world: Making use of cyberinfrastructure. Transactions on data privacy 2008;2-16.
6. Shapiro S. Inside risks - Privacy by design: Moving from art to practice. Comm. of the ACM 2010;**53**:6.
7. Weitzner DJ, Abelson H, Berners-Lee T, et al. Information accountability. Comm. of the ACM 2008;51:82-7.
8. O'Hara K, Shadbolt N. Privacy on the Data Web. Comm. of the ACM 2010;**53**:39-41.
9. Sapuntzakis CP, Brumley D, Chandra R, et al. Virtual Appliances for Deploying and Maintaining Software. LISA USENIX 2003;181-94.
10. Kum, H.C., Ahalt, S, Pathak, D. Privacy Preserving Data Integration Using Decoupled Data. Security and Privacy in Social Network, by Y. Elovici, Y. Altshuler, A. Cremers, N. Aharony, A. Pentland (Eds), Springer 2012;In print.
11. Fienberg SE. Confidentiality, privacy and disclosure limitation, Encyclopedia of Social Measurement, Academic Press 2005;**1**:463-9.
12. Nissenbaum HF. Privacy as Contextual Integrity. Washington Law Review 2004;**79(1)**:19-158.
13. King G. Ensuring the data-rich future of the social sciences. Science 2011;**331**:719-21.
14. McSherry F. Privacy Integrated Queries. Proc. of the ACM SIGMOD 2009.
15. Adam NR, Wortmann JC. Security-control methods for statistical databases: A comparative study. ACM Comput. Surv 1989;**21(4)**:515–56.
16. Mirkovic J. Privacy-safe network trace sharing via secure queries. NDA 2008.
17. Malin B, Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. Jamia 2011;**18(1)**:3-10.
18. Dwork C. Differential privacy: A survey of results. Theory and Applications of Models of Computation. LNCS 2008; **4978**:1-19.
19. Ciriani V, Vimercati S, Foresti S, et al. k-Anonymity. Advances in Information Security 2007.
20. Emam K, Dankar FK. Protecting Privacy Using k- Anonymity. J Am Med Inform Assoc 2008;**15(5)**:627–37.
21. 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.
22. Krenzke T, Hubble D. Toward Quantifying Disclosure Risk for Area-Level Tables When Public Microdata Exists. Section on Survey Research Methods. JSM 2009.
23. Owen P, Shoffner M, Wang X, et al. Report on SMRW, TR-11-01, Secure Medical Research Workspace 2011.
24. Lane J, Schur C. Balancing access to health data and privacy: A review of the issues and approaches for the future. Health Service Res. 2010;1456-67.
25. Jahanian F. Virtual Center for Network and Security Data. University of Michigan IRB 2011.
26. Center for Disease Control and Prevention (CDC), NCHS Research Data Center (RDC). http://www.cdc.gov/rdc/
27. U.S. Census Bureau, CES Research Data Center (RDC). http://www.census.gov/ces/rdcresearch/index.html

28. 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.
29. Boscoe FP, Schrag D, Chen K, et al. Building capacity to assess cancer care in the Medicaid population in New York State. Health Services Research 2011;**46(3)**: 805-20.
30. Trottini M, Fienberg SE, Makov UE, et al. Additive noise and multiplicative bias as disclosure limitation techniques for continuous microdata: A simulation study. J. Comp. Methods in Sci. and Eng. 2004;**4**:5-16.
31. Duncan G T, Keller-McNulty SA, Stokes SL. Disclosure risk vs. data utility: The R--U confidentiality map. NISS TR-121 2001.
32. Ting D, Fienberg SE, Trottini M. Random orthogonal matrix masking methodology for microdata release. IJICS 2008;**2**:86-105.
33. Lahiri P, Larsen M. Regression analysis with linked data. J. Amer. Statistical Assoc. 2005;**100(469)**:222-230.
34. Baldi I, Ponti A, Zanetti R, et al. The impact of record- linakge bias in the Cox model. Journal of Eval. in Clinical Prac. 2010;**16**:92-6.
35. Scheuren F, Winkler W. Regression Analysis of data files that are computer matched, Part II. Survey Meth 1997;**23**:157-65.

APPENDIX : DETAILS OF THE RISK AND UTILITY ANALYSIS

Figure 5 is the scenario for doing the full analysis in the two different settings. As seen in the figure, the scientist has direct access (HIGHER UTILITY) to much more data in the proposed model leading to three important ways in which the utility of the data is increased. First, record linkage is carried out on the restricted access computers by the scientist, who can control and carry over the measurement error in record linkage to the survival analysis (UTILITY 1). Recent literature in statistics have shown the importance of accounting for the linkage errors during the analysis of merged data [33, 34, 35]. Second, the researchers are able to directly run cross tabulations and frequencies on the full data in the restricted access computers when selecting the final attributes and sample. Similarly, if the initial research design needs to be adjusted for additional variables, the researchers can easily reconstruct the customized file from the restricted access system rather than having to go back to the collaborators (UTILITY 2). In the conventional setting, both of these activities are carried out by the collaborator at the health department, with researchers only having indirect access through the collaborator giving them limited control. Finally, researchers can utilize the probabilities of the record linkage in the models for survival analysis (UTILITY 3). Using breast cancer data, Baldi et al. demonstrated that survival analysis on merged records that do not consider issues in record linkage can lead to inefficient and biased results [34]. Better access to data and training that can allow researchers to control the errors during the linkage process will be important in analyzing merged medical records in the future.

Along with better utilization of data, the proposed model reduces risk of disclosure of confidential information in three important ways. First, in the conventional setting of monitored access the risk of data leakage by authorized users is high (RISK 1). There is typically no mechanism for blocking users from taking the confidential data off the system, bringing new data onto the system for linkage attacks, or running sophisticated cryptographic attacks on the de-identified data. The system relies solely on compliance by the user. In contrast, on the controlled access system the researcher is using a locked down VM with all data channels blocked from input and output. Thus, similar attacks will require significant effort to break into the VM raising the cost of attack (REDUCED RISK 1). In addition, malware on a PC can potentially attack data on a monitored access system by manipulating the secure connection software that is used to communicate with the secure server remotely (RISK2). In the proposed model, remote access relies on a locked down VM. It only uses the hardware of the compromised host OS which cannot infect the server (REDUCED RISK 2). Finally, in the health department system, a fully identified table of all merged records exists before it is de-identified, opening

up opportunities for an insider attack by internal staff (RISK 3). In comparison, in the restricted access decoupled data system, after the original data from each of the sources (CR, Medicaid, and death records) is ingested, all data is always maintained as decoupled data. This means that all connection information from the sensitive data to the PII is kept encrypted. Since the connection information from the sensitive data to the PII is never needed outside the computer system, it will never be revealed to a person. Thus, as the record linkage is carried out, the merged table comes together as a decoupled table to start. No identifiable merged table is ever created on any system fundamentally eliminating the potential of insider attack (REDUCED RISK 3). Recognizing the potential for harm in PII data, the proposed model restricts and monitors all activities on and off the computer including print outs never leaving the facility (REDUCED RISK 4).

Finally, we note that typically even de-identified health data used for research tends to include sensitive data such as full dates of service. Under the safe harbor rules, full dates are considered sensitive information which increases risk in data but accurate dates are also important attributes for most analysis. We believe that more data can be safely analyzed in both the restricted access and controlled access systems which are locked down computer systems compared to the conventional monitored access system. In our running example, researchers at the university probably had access to the full dates of birth, diagnosis date, and Medicaid enrollment date so that they could calculate the required variables for analysis such as categories for age at diagnosis. Furthermore, researchers would need to freely conduct sensitivity analysis around timing of Medicaid enrollment and diagnosis of cancer to define the best cutoff of 6 months. These data would have lower risk on the controlled access system compared to the monitored access systems.

We now extend the example to a hypothetical future research on this data to evaluate the two stages in the model for using aggregate data. For this research, we assume that we want to understand the higher risk of death from HL patients on Medicaid for different county of residence. We also assume that we have linked county of residence information in the first step from an available source like Medicaid enrollment. There are 62 counties in NY state with only two counties that have population under the 20,000 cutoff specified for geographic information in the Safe Harbor standard. Hamilton county has about 5000, and Schuyler is just short of 20,000. This means that under HIPAA, when releasing microdata at the person level, the county of residence for individuals living in these two counties would have to be combined as one region of {Hamilton or Schuyler} before being released as de-identified data to a monitored access system. Thus, de-identified data at this stage will not be able to do analysis for these two counties accurately, but rather have one row representing cancer treatment in the combined region. But Hamilton is in the Adirondack region and Schuyler is in the southern tier, and they have very little in common. Thus, there is little reason to study these counties together, making the combined information meaningless. The other option is to combine the small county information to a neighboring county with more meaning, such as Hamilton with Essex, but that results in loss of additional information about the neighboring two counties resulting in a total of four counties whose information is lost. The difference between monitored access to aggregate data and sanitized aggregate data is in the utility of the data. In the monitored access system, at the cost of being explicit with the research activity by filing an exempt IRB and getting an authorized login so that the research activity can be monitored, the researchers can get access to information about all 62 counties with no sanitation. In comparison, in order to build a sanitized dataset for full release to the public, data about Hamilton and Schuyler might need to be sanitized in some way.