Testimony of Bradley Malin, Ph.D.
Assistant Professor of Biomedical Informatics
Vanderbilt University Medical Center

Before the U.S. Department of Health and Human Services AHIC
Confidentiality, Privacy, and Security Workgroup

June 22, 2007

Good afternoon. My name is Bradley Malin and I am an Assistant Professor of Biomedical Informatics at Vanderbilt University. Before providing my testimony, I would like to thank the Chair, Director, and Members of the AHIC CPS Workgroup for the opportunity to participate as part of this afternoon's panel. Today, I will speak with you about privacy issues that are emerging in the integration of genomic and electronic medical records for personalized healthcare research.

I am currently faculty at the Vanderbilt University Medical Center (VUMC) in Nashville, TN, where I conduct research in medical informatics with a focus on data privacy issues that are inherent to the collection, storage, and sharing of patient-specific electronic health information. For years, I have investigated how seemingly anonymous health records, such as DNA sequences, can be "re-identified" to named individuals. In this testimony, I hope to draw upon examples of my research to illustrate that unanticipated re-identification is not only possible, but poses a significant threat, for emerging health information environments. Though the threat of re-identification is real, I will provide examples of how privacy enhancements can be integrated into these environments without limiting the flow of information to scientists who conduct personalized healthcare research projects.

As AHIC has recognized with the establishment of the Personalized Healthcare Workgroup, the practice of medicine is evolving towards personalization. The increasing integration of information and high-throughput technologies into healthcare environments has enabled the collection of detailed genomic and clinical records. In turn, scientists have access to unique datasets to assist in efforts to personalize diagnostics, treatments, and healthcare services. However, the quantity of data necessary to conduct the research that leads to personalized care is often beyond the capabilities of an individual researcher or institution. Thus, it is necessary for researchers to share information collections on a larger scale. Health information exchanges provide an opportunity to share patient-specific records for population-based research projects across organizational boundaries, but at the same time the sharing of such information for purposes other than direct patient care raises significant concerns regarding patients' privacy rights. To enable data sharing for biomedical research, it is crucial that patient-specific data is shared in a manner that protects the identities of the patients.

To protect a patient's identity, it is necessary to understand what makes a record "identifiable". In the context of HIPAA, the Privacy Rule enumerates eighteen elements that are considered potentially identifying features. These elements include names, Social Security Numbers, phone numbers, and various demographics. Note, HIPAA does not explicitly designate a patient's genomic data as a personal identifier. As such, privacy protections for genomic information could arguably be satisfied through de-identification according to the Safe Harbor policy of the HIPAA Privacy Rule, which

requires the removal of the aforementioned elements. Following this reasoning, various computational techniques have been developed and deployed to automatically de-identify an individual's genomic data. De-identified biological data appears protected because there is no public directory that maps genomes to named individuals. Yet, de-identified DNA records, even those that adhere to Safe Harbor can be re-identified to named subjects. This is cause for concern because re-identification does not require "hacking" into secured computers. Instead, significant quantities of seemingly anonymous records can be linked to personal names in publicly available resources through simple automated methods.

Why are de-identified genomic records susceptible to identity compromise? Re-identification of genomic data occurs when the following conditions are satisfied:

1) the data is unique, and

2) the data can be linked to identified records.

The first condition is achieved when unique values reside in the shared biomedical records. Data uniqueness is important because it means that we can distinguish between subjects' records in a shared collection. DNA uniqueness is relatively easy to satisfy and it is estimated that less than 100 single nucleotide polymorphisms, features common to genome-based studies, can uniquely represent an individual. Data used for personalized healthcare research has an even greater potential to be unique because it will supplement genomic records with various clinical, lifestyle, and pharmacological information.

Though uniqueness is a necessary condition it is insufficient to claim re-identification of genomic data will occur. To complete a re-identification, we need a mechanism to link de-identified data to a record that reveals the identity of the subject. This begs the question: Where do we find identified information that relates to genomic records? The answer to this question is heavily dependent on how the genomic records are shared and what it reveals about an individual. For instance, many genomic records are accompanied by genealogical information, often in the form of pedigrees, which assist in family-based population studies. Though the shared pedigrees and genomic records are de-identified, family relations of named people can be reconstructed from public records, many of which can be automatically extracted from the Internet. Recently, I built a software program that extracted genealogical knowledge on current populations from online newspaper obituaries, which report the name of the deceased and, in many instances, the names of the deceased person's relatives. Evaluation of the program with genealogies extracted from a particular U.S. state capital demonstrated that a majority of the current population was identifiable.

Genealogical information is one route by which genomic records can be re-identified, but such data is not always disclosed or available in the public realm. Nonetheless, genomic data can be re-identified by many other routes. As a second example, consider that certain types of publicly available health data collections, such as de-identified hospital discharge summary databases, have been shown to be re-identifiable to public resources, such as voter registration lists. Beyond the reporting of summarized clinical results, discharge records can reveal DNA-specific features, such as a specific mutated gene; e.g.,

cystic fibrosis or Huntington's disease. Similarly, ethnic and gender-specific features can be derived from genomic records using ancestry informative approaches. To leverage such relationships, Latanya Sweeney (faculty at Carnegie Mellon University) and I designed software that combined existing biomedical knowledge to link DNA records to identified discharge records. We extracted eight populations diagnosed with gene-based diseases from hospital discharge databases and our experiments revealed that almost all patients were re-identifiable.

These two examples illustrate that the fallibility of de-identification stems from inferences and features that can be extracted from a shared record. Yet, even when a record appears to lack insufficient inferences, re-identification problems persist. This is because data protection policies are often designed to address an organization's health records without regard to other organizations' data collections. Alone, each policy is sound; however, the protections afforded by the policies can erode when multiple organizations' data collections are brought together. As a third example and an illustration of this problem, consider the following scenario. To protect patients' privacy, a data holder discloses de-identified DNA records to a health information exchange. Similarly, the data holder discloses a collection of identified data, devoid of DNA, for administrative or quality control purposes, such as hospital discharge reporting. When no inferences exist between the shared databases, the separation of DNA and identity protects privacy, but in decentralized healthcare environments a patient generates similar, and often the same, piece of data to multiple organizations. As a result, a patient's location-visit pattern, or "trail", can be extracted from the set of disclosed databases. The

trail can be observed in the shared databases of sensitive, as well as the identified, data; and the uniqueness of an individual's trail often associates seemingly anonymous data with the name of the individual from whom it was derived. Evaluations with the aforementioned discharge populations revealed that significant portions of patient populations are vulnerable to this attack on privacy. This type of re-identification is of significant concern for emerging regional health information exchanges in which disparate data providers collect information on overlapping populations.

Up to this point, my testimony has concentrated on ways in which genomic data is susceptible to re-identification in emerging information exchanges. To an extent, re-identification is possible because the Safe Harbor and Limited Data Set specifications of the HIPAA Privacy Rule do not provide an indication of the identifiability of health data. However, vulnerability does not imply that we can not mitigate re-identification threats. In fact, the necessary conditions for re-identification, uniqueness and linkage to identifying data, provide two clear points at which we can control data identifiability. Specifically, we can 1) prevent the linkage to identifiable information; or 2) prevent the uniqueness of data shared for research purposes. Regardless of the point at which protection is employed, it is crucial that protections are designed and implemented with a formal basis.

What does it mean to formally prevent linkage? When we know what makes data re-identifiable, we can augment the data to prevent the linkage route. As an example, recall the trail re-identification problem. In this scenario, data holders can not prevent the

dissemination of identified information for quality assurance purposes. However, they can collaborate to determine DNA records are re-identifiable by their trails *before* the data is shared. Recently, I devised a computer program that provides data holders with the ability to determine which data holders should suppress information from their disclosed databases in order to guarantee that trails can not be uniquely matched to reveal a DNA records corresponding identity. Moreover, the program allows data holders to calibrate the level of protection that balances policy requirements and a scientist's needs for research support. For example, imagine that data holders come to a consensus and agree upon the following policy: each DNA record should be linkable to no less than 50 patients in a population. After suppressing information to reach this level of protection, there are a certain number of records, say 500, that are available for research. If a researcher needs additional records to conduct a hypothesis test regarding the personalization of health services, say 1000, then the administrators of the data can relax their protections, such that each record links to say 10 patients, and amend oversight (possibly through increased auditing) for the researcher as deemed appropriate. Thus, we integrate policy with formal technical controls over identifiability.

The previous example is an illustration of how formal privacy protection models can be integrated into electronic health information exchanges. Yet, it is only one such example of privacy. In reality, what is needed is a quantification of re-identification risk associated with each record that is shared with an information exchange. As the quantity of re-identification risk for shared records increases, so too must the oversight for such

data used in research studies. Similarly, as the re-identification risk decreases, the oversight can be relaxed.

One point that is worth remembering: identifying information that is available for re-identification purposes is not regulated by the confines of health information exchanges. Rather, this is information that is disclosed for other purposes and will not be removed from the public realm. Thus, data protection for emerging health information environments, must be cognizant of the information that already exists in the public realm. As such, it is necessary for entities disclosing information to health information exchanges to consider the information that exists beyond their own organizations prior to disclosure.

Finally, when healthcare organizations believe that patient anonymity is at risk, they should complement technological protections with contracts, such as a data use agreement, in which the data recipient pledges not to attempt re-identification of the subjects. Contractual agreements provide a legal basis that, in fear of heavy fines and the potential for imprisonment, discourages improper use of patient data. Contracts and legal agreements do not diminish the fact that the data is susceptible to re-identification, but in combination, technology and policy can provide clear and enforceable oversight.

I thank you for your time and dedication to this topic.