# Repurposing Clinical Data for Cancer Research With Formal Privacy Protections
# President's Cancer Panel
# December 14, 2010
# Bethesda, Maryland

Bradley Malin, Ph.D.
Assistant Professor
Department of Biomedical Informatics, School of Medicine
Vanderbilt University
Nashville, Tennessee

First and foremost, I thank the Panel for the opportunity to testify on recent developments in health information technologies and the opportunities, and potential challenges for, cancer research.

## A Movement Toward Clinical Information Reuse

The healthcare community is in the midst of an information technology (IT) revolution. Over the past several decades, advances in health IT, such as electronic medical record (EMR) systems, have facilitated the collection of large quantities of finely detailed personal data. At the same time, the field of personalized medicine has been fueled by the aid of a number of ventures, such as the Personalized Health Care Initiative of the U.S. Department of Health and Human Services. In turn, there has been a magnification in the desire to utilize information embedded in EMRs to tailor a patient's preventative care and treatment regimen to their specific evidence, which is crucial for the management of cancer. Given the detail of information in EMRs, and their increasing integration with biorepositories, EMRs are poised to become a critical component of biomedical research endeavors.

More importantly, EMR systems can amass data on sizable populations, which can be automatically mined for a substantial range of clinical phenomena, on a massive scale. The opportunities are awesome, and demonstration projects, such as those undertaken by the NIH-sponsored Electronic Medical Record and Genomics Network (eMERGE) consortium, have shown that EMRs can be repurposed for large-scale clinical phenotyping and subsequent genome-wide association studies (GWAS). Though EMRs contain noise and error, eMERGE investigations indicate that highly-specific clinical phenotype models can be specified for such systems in a manner that is reusable across diverse systems. Already, models for over ten clinical phenotypes have been developed and utilized in GWAS with over 17,000 patients. The EMERGE network currently consists of five primary sites, including the Group Health Cooperative of Puget Sound, Marshfield Clinic, Mayo Clinic, Northwestern University, and Vanderbilt University, and a recent RFA from the NIH suggests that eMERGE will expand in size. Moreover, other initiatives, such as the Pharmacogenomics Research Network, are broadening in scope and reuse of EMR systems as well. Much more could be said about how phenotyping from EMRs can be achieved for cancer research purposes, but the goal of this testimony is to focus on the privacy issues, and so I refer you to the eMERGE website for further references on this topic (http://www.gwas.net).

## There is a Tension Between Clinical Data Sharing and Patient Privacy

Until recently, the collection, analysis, and application of clinical and genomic information were localized to specific investigators or institutions. Increasingly, however, sharing data beyond the borders of the initial collecting organization has become a vital component of emerging biomedical research frameworks. Clinical scientists need to share data to strengthen the statistical power of complex association experiments, allow the research community the opportunity to replicate and verify clinically-relevant findings, and comply with a host of regulations. In fact, it is of such importance that in the United States, certain federal agencies, including the NIH, have adopted policies that mandate sharing data generated, or studied, with federal funding. To facilitate such actions, agencies around the globe have invested considerable effort to construct IT infrastructure, such as the Database of Genotypes and

Phenotypes at the NIH, which will assist in the consolidation, standardization, and dissemination of participant-specific records from disparate investigators.

At the same time, the increased collection and sharing of sensitive biomedical information has raised significant societal issues, including concerns over patient privacy, which could easily derail these efforts. To attempt to resolve such conflicts, various policies and regulations embed privacy respecting safeguards. In particular, regulations tend to permit sharing of health information without patient authorization provided that data is "de-identified", which could help to mitigate bias in large-scale cancer research studies. Even when healthcare providers seek the consent of their constituents before sharing data, such consent is often tied to the promise that the data will be shared in a non-identifying manner.

However, this begs the question of what exactly is de-identification? In the U.S., the Privacy Rule of the Health Insurance Portability Accountability Act (HIPAA) is general and states "*Standard: de-identification of protected health information.* Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information". Technically, to achieve this goal, the Privacy Rule delineates several routes by which data can be rendered de-identified: 1) Expert Determination and 2) Safe Harbor. The Expert Determination approach requires an expert to certify that "the risk is very small that the information could be used….to identify an individual who is the subject of the information [using] generally accepted…methods for rendering information not individually identifiable". The regulation points to various methods applied by federal statistical agencies, and some candidates for supporting this approach for health information in particular were suggested in the comments to the legislation. An alternative route specified by the Privacy Rule is Safe Harbor, which enumerates eighteen attributes that must be suppressed. The Safe Harbor policy was designed to be "an easily followed cookbook approach" for cases where the certification required by the statistical standard was too onerous; it is fundamentally the mechanism that the NIH Data Sharing Policy points data managers to for guidance when constructing their data sharing plans.

In practice, most healthcare organizations shy away from the expert standard in favor of Safe Harbor. This is probably not because it is a preferred option, but is likely due to various factors, including: 1) there are no standardized methods (or consensus) for satisfying the expert approach within the HIPAA Privacy Rule, 2) there is a lack of readily available open source software for applying methods that mitigate re-identification risk in health information, and 3) health managers often find it difficult to determine the identifiability of health information in practice.

Yet, it is important to recognize that de-identifying data according to the Safe Harbor standard is less than ideal for cancer researchers because it limits the ability to perform various studies. For instance, cancer epidemiologists often require detailed geographic information; however, Safe Harbor only permits the disclosure of the first three digits of a ZIP code, which can limit accuracy in model development and evaluation (Boulos 2009). Similarly, cancer researchers that focus on the elderly are hampered by the fact that Safe Harbor requires all ages above 89 to be forced into a single top-coded value of "90 or greater".

## Re-identification Risks are Contextual

Perhaps an even greater concern is that the Safe Harbor policy does not eliminate the risk of "re-identification" of a shared record to the name of the patient from whom it was derived. For instance, in testimony by Latanya Sweeney before NCVHS (Sweeney 2007), it was shown that, according to U.S. Census statistics, a certain portion of the U.S. population is expected to be unique based on the demographics left behind by Safe Harbor. This is a concern because such demographics can be found in publicly available resources, such as voter registration lists (which were used in a high profile re-identification of the Governor of the State of Massachusetts in the 1990's). And, in a more recent study we showed this result is highly variable - dependent on the U.S. state from which the data is shared (Benitez 2010b). The main point: risk to re-identification of data that adheres to Safe Harbor is not zero.

Additionally, when we consider data derived from EMRs for cancer research purposes, it should be noted that re-identification of seemingly Safe Harbor-compliant health information is not limited to residual demographics. For instance, a patient's clinical profile can be leveraged for re-identification purposes as well (Loukides 2010). This is a potential concern because clinical phenomena are often disclosed through standardized terminologies; e.g., the International Classification of Diseases (ICD), that are replicated in both de-identified health information as well as in identified resources, such as the EMR from the data was derived or a hospital discharge database. To determine the magnitude of such vulnerability, we performed a study with approximately 3000 patient records with which Vanderbilt researchers recently performed a GWAS on native electrical conduction within the ventricles of the heart. Our investigation indicated that the combination of diagnosis codes within 96% of these records made them unique - even in context of the 1.2 million patient population from which they were derived.

At this point, it is important for the Panel to recognize that the existence of a re-identification route for health information is not the same as exploiting the route. Again, this is because re-identification risks are contextual. For instance, in the voter registration attack, it was estimated that only 0.04% of the U.S. was expected to be unique based on the combination of {*Year Birth*, *Gender*, 3-*digit ZIP*} (Sweeney 2007). Moreover, each U.S. state has different guidelines or regulations on release of public records. Our investigations showed that availability and monetary cost to purchase and utilize such resources was variable and, in many instances, cost prohibitive (Benitez & Malin 2010). Consider, in the state of South Carolina, we estimated that approximately 1,386 records Safe Harbor compliant records could be re-identified at a cost of $0 (where the voter registration list is free), but in the State of Wisconsin, where the Marshfield clinic resides, we estimate that only 2 records could be re-identified at a cost of $6,250 each.

It is possible that a level of risk is necessary to develop ethically sound research policy and avoid stagnancy in biomedical research, but without understanding the magnitude of risk, it is virtually impossible to make informed decisions. Knowledge of the actual re-identification risk associated with a given dataset would help resolve if data is underprotected and in need of additional safeguards, or overprotected such that data sharing policies could be more permissive.

**We Can Share Clinical Data with Formal Guarantees about Privacy**

In this testimony, I advocate for de-identification frameworks that are more informed about re-identification risks and leverage the Expert Determination standard. Again, our recent work with the eMERGE consortium illustrates how decision makers can model and measure privacy risks in an easy to digest manner with respect to existing policies in the context of known threats.

To date, various approaches have been proposed to support the Expert Determination process. For instance, prior to our research, there were models alluded to in the HIPAA Privacy that were designed to determine the amount of re-identification risk in a particular record or set of records. In the statistics community, common models have focused on estimating how many records correspond to unique people. Other models have taken a more broad perspective and model how many individuals in the population a record could have been derived from. Such approaches then set a threshold on the probability of identifying a record to a specific individual or set of individuals, considering records at risk if this threshold is exceeded. This is similar to a model studied in the computer science community, known as *k*-map, which states that each record is protected when it corresponds to *k* people in the population from which it was derived (Sweeney 2002). Under this model, it can be guaranteed that the likelihood of identifying any particular record is at most 1/*k*.

In certain instances, population data is readily available. For instance, an expert could leverage public statistics from the U.S. Census Bureau to estimate the size of the population group for each patient record to be shared (e.g., the number of 52-year old Caucasian males in ZIP 372**). However, there are times when a health data manager may not know the details of the population; for this case, a more strict model called *k*-anonymity has been proposed. Technically, this model is satisfied when each shared record is equivalent to *k*-1 other shared records. This model, also referred to as a *binning* strategy, has received increasing attention in the healthcare domain over the past several years (See El Emam 2008).

Binning models were originally designed for relational data, such as demographics, and were recently extended to account for variable length transactional data, such as the clinical profiles (e.g., combinations of diagnoses) alluded to earlier. Yet, our investigations revealed that algorithms to support the latter, which were developed within the domain of computer science, were ill-suited to support clinical

**Table 1.** Ability for diagnosis anonymization strategies to retain sufficient clinical information for research purposes.

| Disease | Anonymization Approach | |
|---|---|---|
| | Vanderbilt | Competing |
| Asthma | ✓ | ✗ |
| Attention Deficit Disorder with Hyperactivity | ✗ | ✗ |
| Bipolar Disorder –Type 1 | ✓ | ✗ |
| **Bladder Cancer** | ✗ | ✗ |
| **Breast Cancer** | ✓ | ✗ |
| Coronary Disease | ✓ | ✓ |
| Dental Caries | ✓ | ✗ |
| Diabetes Mellitus – Type 1 | ✓ | ✗ |
| Diabetes Mellitus – Type 2 | ✓ | ✓ |
| **Lung Cancer** | ✓ | ✗ |
| **Pancreatic Cancer** | ✓ | ✗ |
| Platelet Phenotypes | ✗ | ✗ |
| Preterm Birth | ✓ | ✗ |
| **Prostate Cancer** | ✓ | ✗ |
| Psoriasis | ✓ | ✗ |
| **Renal Cancer** | ✓ | ✗ |
| Schizophrenia | ✓ | ✗ |
| Sickle-Cell Disease | ✓ | ✗ |

investigations. In particular, this was because they were not cognizant of the manner in which clinical information is coded or the needs for the clinical research community. To overcome such deficits, we developed an algorithm to anonymize transactional data with significantly less information loss than existing methods. To keep the process goal-oriented, we developed an intelligent constraint-based algorithm, which was oriented to maintain diagnoses that are known to be useful in for GWAS studies, including many cancer-related investigations. We recently demonstrated our algorithm on the same Vanderbilt GWAS dataset mentioned above (Loukides 2010). In the process, we compared our method to a competing method published by the computer science community and when we set the value of *k* equal to 5 (a typical value used for protection by various statistical and public health agencies), we found we were able retain a significantly greater amount of the clinical information that would be useful for cancer research purposes. As an example, Table 1 provides a summary of the cancer related diagnoses we were able to retain in comparison to a state of the art anonymization model published by the computer science community. Our approach consistently outperformed, particularly for cancer-related diagnoses. For instance, our algorithm could retain information with respect to breast, lung, pancreatic, prostate, and renal cancers, whereas the competing algorithm was unable to (Loukides 2010).

### Allowing For a Range of Risks When Sharing Clinical Data

Though formally computable, the aforementioned models have several drawbacks if we want to relate the results to the more popular Safe Harbor de-identification policy. In particular, they assume that each shared record should be equally risky. In other words, each record must correspond to a group of *k* or more people. However, the application of such a *k*-based model would preclude de-identification solutions where one record is more vulnerable to re-identification while the rest of the records are comparatively less risky; e.g., solutions like Safe Harbor which leave records in various group sizes, some of which may be unique (i.e., *k*=1). Solutions produced by *k*-based models may therefore require more information loss. Whether this tradeoff is appropriate is best left to the discretion of administrators, but we wish to provide a framework that is amenable to alternative de-identification solutions.

We recently developed such an approach for patients' demographics (Benitez 2010a). The process consists of three steps. First, it transforms patient demographics into their Safe Harbor permissible form. Second, given these demographics, it estimates the re-identification risk of the dataset. Third, it executes a risk mitigation procedure that can be likened to tuning a set of knobs for the fidelity of each demographic attribute. We can dial fidelity down and coarsen an attribute (e.g., the age to five-year age range). Or, we can dial fidelity up (e.g., U.S. state to five-digit zip code). If the risk for the altered cohort is no greater than Safe Harbor then it can be certified as an acceptable solution in accordance with the Expert Determination model. We applied this approach to various eMERGE cohorts, as shown in Table 2, many

**Table 2.** eMERGE cohorts and alternative disclosure policies: A = {*Generalized Ethnicity*}; B = {*Age at 5 Year Intervals*}, C = {*Policy A + B*}, D = {*Age at 10 Year Intervals*}.

| Who | Clinical Phenotype | Cohort | Over 89? | Policy | | | |
|---|---|---|---|---|---|---|---|
| | | | | A | B | C | D |
| Group Health | Dementia | 3,616 | 1,483 | ✗ | ✗ | ✗ | ✓ |
| Marshfield | Cataracts | 2,646 | 269 | ✓ | ✓ | ✓ | ✓ |
| Mayo | Peripheral Arterial Disease | 3,412 | 29 | ✓ | ✓ | ✓ | ✓ |
| Northwestern | Type 2 Diabetes | 3,383 | 6 | ✓ | ✓ | ✓ | ✓ |
| Vanderbilt | QRS Duration | 2,983 | 12 | ✓ | ✓ | ✓ | ✓ |

of which have a significant number of records with “too old” ages; i.e., over 89 (Malin 2011). And, in most instances our approach could find alternative disclosure policies that had an acceptable level of risk.

**How Can We Embed Data Privacy into Future Cancer Research?**

Though we can design privacy enhancing technologies to facilitate the dissemination of patient information for clinical research purposes, there are still challenges to overcome before they are accepted for everyday use in cancer research. I would like to take a moment to highlight several of the challenges.

***What are practical “adversarial” models for cancer research?*** As was alluded to earlier, there are various routes by which health information can be re-identified, but not all are equally likely. The HIPAA Privacy Rule states that data must be protected from a *reasonable* recipient, but there needs to be clarification on what this means with respect to the known resources available and routes by which re-identification attacks can be perpetrated. Moreover, this is all the more important to clarify for the cancer research community because the management and modeling of cancer in a clinical setting is inherently temporal (e.g., treatment-response). Yet the inclusion of temporal information (e.g., time between visits) adds an additional dimension to clinical information that could lead to yet another route for re-identification. It is likely that we can develop privacy protection algorithms to facilitate longitudinal data sharing without violating privacy requirements, but are such algorithms necessary to sufficiently protect against anticipated re-identification concerns?

***Should IRBs make identifiability decisions or do we need to certify de-identification “experts”?*** HIPAA designates that an expert needs to certify the identifiability of clinical data. However, there is currently no certification or training program that trains such experts. At the present moment, the determination of whether a dataset is de-identified tends to falls to an investigator’s IRB. However, it is probably not practical to assume that, at the present moment, the constituents of an IRB have sufficient training to make determinations about identifiability. There are several possible ways in which this expertise could be built up or leveraged. First, a federal agency such as the NIH, or HHS, could standardize and publish educational documents for IRB members on what constitutes de-identification and how to adjudicate if it has been satisfied. Alternatively, a credentialing program could be established by an independent organization that trains and certifies experts. In this way, the IRB could contract with, or retain, experts as necessary. Finally, in the event that no such process can be established, then the NCI may consider partnering with other NIH institutes to establish a Center(s) of Excellence for Data Privacy to ensure that information collected from EMRs is sufficiently de-identified.

**References:**

(Boulos 2009) Boulos M, Curtis A, and AbdelMalik P. Musings on privacy issues in health research involving disaggregate geographic data about individuals. *International Journal of Health Geographics*. 2009; 8: 46.

(Benitez 2010a) Benitez K, Loukides G, and Malin B. Beyond safe harbor: automatic discovery of health information de-identification policy alternatives. *Proc. of the ACM International Health Informatics Symposium*. 2010: 163-172.

(Benitez 2010b) Benitez K and Malin B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*. 2010: 17(2): 169-177.

(El Emam 2008) El Emam K and Dakar F. Protecting privacy using *k*-anonymity. *Journal of the American Medical Informatics Association*. 2008; 15(5): 627-637.

(Loukides 2010) Loukides G, Gkoulalas-Divanis A, and Malin B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc. of the National Academy of Sciences*. 2010; 107: 7898-7903.

(Malin 2011) Malin B, Benitez K, and Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association*. Forthcoming.

(Sweeney 2002) Sweeney L. *k*-anonymity a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems.* 2002; 10(5): 557-570.

(Sweeney 2007) Sweeney L. *Testimony before the National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information*. August 23, 2007.