

Building a Chain of Trust: Using Policy and Practice to Enhance Trustworthy Clinical Data Discovery and Sharing

Nick Anderson, Ph.D.

Department of Medical Education and Biomedical Informatics
University of Washington
Seattle, WA 98109
1-206-685-0249
nicka@uw.edu

Kelly Edwards, Ph.D.

Department of Bioethics and Humanities
University of Washington
Seattle, WA 98195
1-206-221-6622
edwards@uw.edu

ABSTRACT

Advances and significant national infrastructure investment into clinical information systems are spurring a demand for secondary use and sharing of clinical and genetic data for translational research. In this paper, we describe the need for technically leveraged policy models and governance strategies to support data sharing between a range of disparate stakeholders where trust is not easily established or maintained.

Categories and Subject Descriptors

E.4 [Data]: Coding and information theory – *formal modes of communication*

General Terms

Security, Theory, Legal Aspects, Verification

Keywords

Policy governance, compliance with government regulations, trust, data sharing, clinical data, translational health research

1. INTRODUCTION

Large-scale national initiatives such as the NIH/NCRR CTSA translational roadmap [1], the National Center for Biomedical Computing (NCBC) sites, the NCI caBIG consortia, and the Office of the National Coordinator for Health Information Technology (ONC-HIT) are collectively stimulating a common level of technical expertise, evolving resource infrastructure and motivations to discover, share, request and analyze clinical and clinically acquired genetic data for research. This demand and the corresponding technical capabilities are certain to grow, and with it grow the challenges to facilitate access to and uses of such data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GTIP 2010 Dec. 7, 2010, Austin, Texas USA
Copyright 2010 ACM 978-1-4503-0446-7/10/12 ...\$10.00.

that protects privacy while advancing health research. The intersection of health information systems and policy is being defined at the national level - the Office of the National Coordinator for Health Information Technology (ONC-HIT)'s first "area for consideration" of the HIT policy committee is "technologies that protect the privacy of health information and promote security in a qualified electronic health record". However, despite these common competencies, opportunities and engaged stakeholders, there is a lack of best practices of how to establish effective data sharing policies that balance assurance of protection to patients, institutions and researchers against research utility and obligation to research and patient communities.

Traditional approaches to gaining approval for using clinically derived data follows a common pattern that can be considered a closed system. A researcher defines a protocol describing the scope and scale of their requirements, and submits this definition along with an application to their Institutional Review Board (IRB) or similar regulatory body. The IRB evaluates this protocol in context with institutional human subjects policy, state and federal law and if it meets local expectations, issues approval and requirements to manage compliance for specific focused research. In this context, there is an established trusted relationship between researcher, subjects and institutions based on explicit description of patient privacy protection, and is supported by a fair examination of potential risks. For the significant majority of such research, IRB approval is typically sought prior to or during the initial phases of a project, and is rarely revisited unless modifications need to be made or if unforeseen consequences arise - such as privacy disclosure events or incidental findings.



Figure 1: Traditional research data sharing model

The approach to upfront review has developed in the context of discrete research projects such as clinical trials, where the stakeholders are limited in number, the intervention is clear and the harms are known or calculable. This (to date) largely functional approach becomes decreasingly useful when the relationships between research stakeholder, institution, patient and secondary intent become more complex or numerous. Where upfront approval that captures the relationship between a researcher and his or her finite number of clinical patient subjects

has been historically relatively straight-forward to establish in advance, it becomes burdensome or impossible to establish when such research expands to managing privacy concerns in population level patient cohorts. As a consequence, up-front IRB approval for many large-scale data-dependent projects – integrated research clinical data repositories, community-wide comparative effectiveness research, genome wide association studies, population health surveys or large scale disease registries – increasingly depends heavily on de-identification of data at the point of acquisition and prior to being made available for analysis as to manage risk of privacy disclosures in what are in practical terms open data sharing systems.

It is an embraced guidance from the Office of Human Subject Research Protection (OHRP) that properly de-identified data are no longer considered human subjects (see US DHHS 45 CFR 46.101(b)(4)). However, this path to de-identification as to minimize risk and gain regulatory approval is increasingly considered to be of only limited effectiveness by itself. In fact, the entire concept of de-identification has been questioned as oversimplified and insufficient in attempts to solve the increasingly complex problems of clinical data use and access that are facing the growing national and global research community. While effective from a regulatory perspective at least when the focus is on quantitative and structured text or meta-data, the practice of sanitizing clinical data to sufficiently render it a “non-human subject” may or may not protect key stakeholders (e.g. patients, clinicians, clinical organizations) and is frequently positioned in philosophical opposition to research utility and public benefit. It is quite easy to reach a point of technical data de-identification that not only greatly limits actual research analysis, but can inadvertently discriminate against entire classes or populations of patients by rendering them “invisible” – that is, protected by removing all potentially identifiable populations from the data sets as to adhere to strict HIPAA guidelines [2,3]. Such protections, while compliant with a strict interpretation of current law, can disproportionately impact the very patient communities that are most in need of modern research, for example, patients from under-represented minority groups, patients with a rare genetic disease, or patients in rural communities with common health disorders.

There are other scientific limitations to de-identified and de-linked datasets in that increasingly our common and complex diseases require richer, thicker, longitudinal data to identify the multi-factorial contributors to disease or survival. Thus, this process of de-identifying patients poses both moral and scientific challenges by turning patients into mere “dreams or dots” [4]. Literally de-humanizing datasets – removing the connection to human subjects – can work against our interests in promoting respectful stewardship of data. For example, we know from behavioral science that the potential to inflict harm increases when the subject is anonymous or unseen by the actor [5]. A central challenge in research practice today is exactly how to put a human face on the data while still protecting individual privacy.

Architectural decisions established when building clinical data discovery systems and sharing data for research are often based in managing perceptions of risk and can be tied to lack of formalized trust relationships between stakeholders. In our traditional research models, trust was given from patients/participants to an individual clinician/investigator. That investigator was then the steward of these data and shared with known, trusted

collaborators. However, the very rapid scale, vast scope and sheer quantity of data sharing for research has changed that intimate trust landscape. As we have seen with other high-profile lapses, the research enterprise as a whole has much at stake in getting these handoffs right [6,7] – and this challenge crosses disciplines and communities. In this paper, we focus on current challenges and strategies for operationalizing this chain of trust as it expands, and suggest future areas of technically leveraged policy development.

2. ESTABLISHING A BASIS FOR TRUST

As outlined above, a trust relationship for clinical research has traditionally been defined and encoded between investigators and participants in a structure that minimizes risk through maximizing human subjects protection through de-identification processes as well as often specific and directed consent agreements between patients and researchers.

Increasingly, there are new considerations to this model, such as how to establish trust relationships between investigators and patient communities as a whole, between institutions on behalf of their patient populations, or between researchers and federal requirements on behalf of their patients. When moving up scales of stakeholders and stakeholder relationships, it becomes difficult to establish who is responsible for certain obligations of data ownership – when data is aggregated or pooled, who is responsible – the originator of a portion of the data, or the data manager? And to whom are the data users accountable? The most common regulatory answer is that downstream data users are held accountable by their home institutions’ policies and IRBs, but from an ethical perspective, it can be argued that there should be some downstream accountability back to original participants.

Just how this chain of trust can be meaningfully passed forward to future users is an open question. We know from past work on trustworthy research practices that building and sustaining trust requires attention to relationships and systems of accountability [8]. In lessons learned from other industries such as airlines or energy, we also know that regulations should just provide the floor for standards of practice; it is up to the research community itself to set standards of excellence that exceed the restrictions set by the regulatory environment which is designed to promote the minimum acceptable risks. How can we assure that systems and processes support the ability for trust obligations to track forward to downstream users? We have Data Use Agreements (DUA) and Material Transfer Agreements (MTA) that again meet our floor regulatory needs, but rarely to-date do we have agreements or knowledge structures that pass along various trust obligations to the original study population (e.g. commitment to conduct research in a certain domain, commitment to return results that are clinically relevant, or commitment to maintain communication about study activities), and less so to the new collaborative models described earlier. In the absence of a professional standard of practice, the research community has primarily followed the regulatory guidance as the best available basis for protecting privacy as well as institutional risk. We review the two most common practices for preserving trust through protecting privacy below.

2.1 De-identification and exemption as a basis for trust

The current OHRP guidance and standards of public health research enforce de-identified data sharing only. We have a long history of using large publicly available datasets for epidemiological studies and other public health projects. This same approach carries over to comparative effectiveness research with clinical datasets, where the individual outcomes do not matter as much as population-level response to different interventions or management strategies. These kinds of established research uses have presumed several things about the public health or health utility value of the research. As members of the public, we are willing to give up certain privacy limitations in exchange for certain benefits, like preventing the spread of infectious disease or tracking and improving medical care. However, this approach to building research resources within institutions is expected to continue and de-identification alone is increasingly being seen to be creating a false sense of security [9]. Gatekeeper roles in the form of data managers or honest brokers take on increasing importance, and there are multiple approaches to combining enhanced de-identification approaches within stewarded environments that are extending investigator abilities [10,11]. Currently, these solutions are typically operating under the same assumptions that trust cannot be easily established for secondary uses and thus focus on protection and secure release of sanitized data.

However, important questions remain which cannot be avoided by further de-identification, such as: Who weighs whether the potential benefits gained through the specific research are worth the trade-offs of potential risks to privacy and other unanticipated wrongs? Does broader data sharing of de-identified data actually accomplish our goals of better translational health research?

2.2 Data aggregation as a basis for trust

Often building on or used in coordination with de-identification approaches, data aggregation or data pooling often provide a similar perceived measure of reducing risk in data sharing environments by de-identifying or obfuscating data sources from the end-users. In many cases, data aggregation occurs in building clinical data resources for research, where it is unlikely or implausible for data providers to establish a relationship with either patients or end users. Aggregation has also been used to support large community public data sets (dbGAB, GenBank, ArrayExpress), or as a mechanism to remove the ability for data providers to be compared or stratified on a 1 to 1 basis (e.g. inter-institutional, inter-repository) - such as in outcomes measures or comparative effectiveness registries. Where aggregation can differ from de-identification approaches alone is when stakeholders seek to minimize risk of end-user data analysis that could lead to either identifiable data, or more subtly, lead to comparisons that may cast the original data providers in a negative light.

Data aggregation approaches reflect a different aspect of defining trusted relationships – that aggregation is perceived to provide sufficient anonymity to data providers who are often geographically separated, at several levels of remove from original data sources, and may be acting on behalf of repositories and or institutions as a whole. With limited control over relationships and end-users, aggregated or pooled and de-identified data is a perceived a necessary and plausible trade-off if the data provider can still gain benefit by being part of a

cooperative, but still has plausible assurance of not being singled out for comparison. Aggregation is currently the basis of a variety of federated data sharing initiatives where sharing of data (whether willingly or mandated) is conditional and is built upon independently de-identified data sources. As in basic data de-identification, unanswered questions remain - such as to what degree does aggregation diminish the ability to measure whether individual data providers are adhering to common data representations that accurately reflect the underlying health information systems? Does an approach to aggregation imply a necessary focus on common lowest common denominator data alignment, and could these common data representations be enhanced with greater understanding and control of how downstream users could use the resulting data resources?

3. SCALES AND MODES OF DATA SHARING

Clinical data discovery and data sharing occur across every conceivable range of scales and stakeholders, and have an associated broad range of risks and opportunities, both perceived and concrete. We work through how the chain of trust currently flows through to downstream uses and relationship through three different research scenarios: between investigator/data owners, between institutions, and finally, through mandated federal data sharing.

3.1 Inter-investigator data sharing

Examples of inter-investigator trust issues are when owners of data resources (such as an investigator maintained biorepository consisting of several hundred biospecimens prospectively collected, phenotyped, and consented for a specific research purpose) are asked to share information about their repository for external discovery. In this context, the original researcher remains the primary data steward and gatekeeper for future uses.

Such key data holders are faced with several challenges. Of primary importance is whether these stakeholders have the capability to share biospecimens that were consented for a specific study with another researcher who may or may not have a formal affiliation. This can sometimes be established by review of the IRB or consent documents, and where necessary can be plausibly addressed by recontacting and reconsenting patients to establish approval for this secondary use [12].

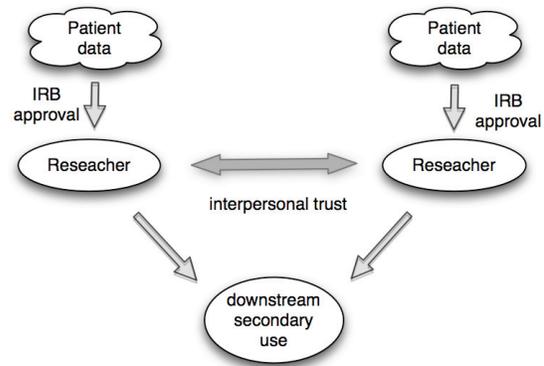


Figure 2: Inter-investigator data sharing model

A different and more human challenge is the individual data owners attitude to data sharing – since most biorepositories to date

have not been established with data sharing in mind – and in fact may have covenants to explicitly not permit this – it is up to the individual stakeholder to evaluate their personal willingness to share information about such valuable research samples, to whom, and then for what purpose. There are also other associated more practical challenges to consider, as to establishing for what benefit and by what resources can data sharing occur – as the typical data manager of a resource such as a biorepository is managing said resource as part of a specific, funded and focused effort – and a data sharing arrangement must at the least be cost-neutral unless other benefit can be gained.

3.1.1 Current Approach to Transferring Trust

The primary method for establishing and transferring trust in this inter-investigator model of data sharing is defining interpersonal trust. Beyond the basic question of whether these stakeholders have the capability to share are the issues of who or what should be accountable for supporting this sharing, and what mechanisms need to be available to support that trust and obligation is transferred to secondary or tertiary users of these specimens? In general, we all know or discover quickly who we want to collaborate with and who is a trustworthy player, and traditionally most such decisions are made on the basis of potential payoff for future collaborations. However, what technical and auditable basis do these primary data holders have to ensure that the two features that will preserve trust - relationships and accountability – will track to the next users, and specifically, how do these define these elements as to manage their own risk?

3.2 Inter-institutional data sharing

With rare exceptions, most large-scale institutional health systems conduct slow and low-scale competitive wars with other institutions in terms of perceived quality and capabilities. With budgets and income of hundreds of millions of dollars annually, research institutions have very real business reasons to not to appear to be providing anything less than the best health care, and are loath to be compared in terms that could affect community perception of their health services. As research uses of these data are increasingly in demand and coupled with new federal requirements to prepare for the sharing of clinical data under ONC HIT Meaningful Use [13], institutions are faced multiple competing challenges to determine what is a manageable risk to their participation in large-scale research data sharing.

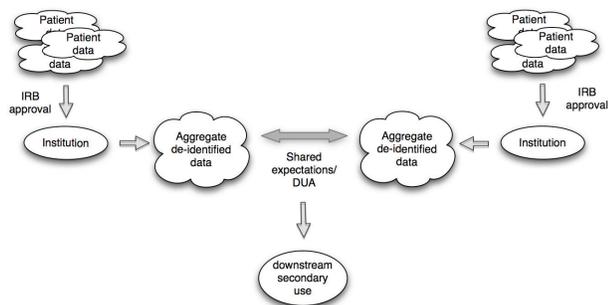


Figure 3: Inter-institutional data sharing model

To date, most inter-institutional research data sharing has been associated with specific funding (and thus largely de novo), occurs between more than 2 institutions or sites that are often geographically separated, and supports the full range of de-identification, obfuscation and aggregation processes described

earlier. Unless the participating institutions have a previously established agreement for management of access to these data resources, the assumption is that each institution must establish their own public view of the data that best minimize individual institutional risk.

The outcome of this form of collaboration can be very large-scale data sets - often in the millions of patients – but which have been established in such a way that it is often impossible to determine original source and which may have been rendered de-identified to the point that effective research analysis is severely hampered [2]. As such, there are challenges to the utility of these approaches that is demanding higher levels of data description that will need to go beyond de-identified aggregate views as to be able to be used to impact health care.

3.2.1 Current Approach to Transferring Trust

Within inter-institutional data sharing, the approach to conferring trust on downstream users (or institutions as “users”) can best be described as a commonwealth of shared expectations. When an institution chooses to enter into an inter-institutional data sharing agreement, it does so with clear, upfront expectations about the purpose of the sharing and the restrictions of such. These expectations are often formalized in a DUA; however, the relationship would not begin without a belief and commitment to the common purpose for the inter-institutional sharing. Like with individual participants, institutions are willing to risk something (potential for exposure) to gain something (participation in a larger network) if there are sufficient trust relationships and protections in place. Establishing these data sharing projects presently requires a commitment from senior stakeholders in clinical, regulatory and research – though it is unlikely that they will remain the steward of such resources once implemented.

3.3 Mandated federal data sharing

The most current example of mandated federal data sharing is the NIH requirement to submit all GWAS data into a federal repository (dbGaP). This requirement stems in part from federal legislation that required all publicly funded projects should be in the public domain. However, problems quickly emerged when the formerly certified de-identified datasets of genotype information were shown by Homer et al. to be identifiable [14]. With that technological possibility in 2008, the genotypic data moved behind a firewall alongside the phenotypic data and requires review and approval from a Data Access Committee (DAC).

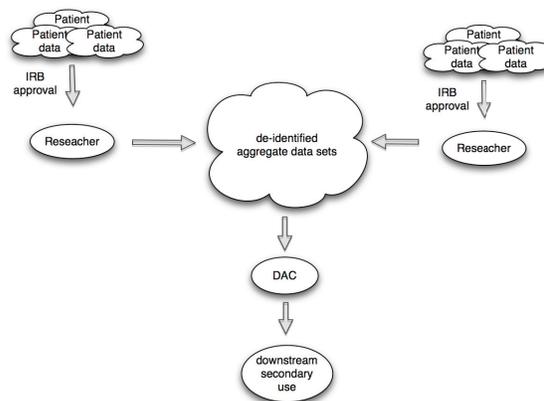


Figure 4: Federal mandated research repository model

The DAC in this case becomes the key steward and gatekeeper to future uses and the primary data holders have little to no control over what happens next to their data. In one instance, such as the 1948 Framingham Heart Study cohort [15], the submitting data holder requested that the secondary data user have an IRB review for their data use [16]. Other datasets are subject only to the local institutional policies of the secondary data user and to any restrictions initially outlined in the original consent form (e.g. limiting downstream uses to schizophrenia research only).

3.3.1 Current Approach to Transferring Trust

In the mandated model, the data sharing is the most anonymous of the designs we are examining. Here we can no longer rely on any measure of interpersonal trust and therefore, the need to operationalize the chain of trust becomes more formal. Downstream users sign a DUA in which they promise to not attempt to identify individuals nor share the data with any additional users (including trainees) who are not on the original DUA. However, there are 13 separate DACs who govern the use of dbGaP data with varying practices around approvals and restrictions. These practices are emerging and are arguably not yet at the level of transparency and predictability that such a system would require. With no interpersonal trust to fall back on, the systems of accountability and auditability need to be even more robust to assure the chain of trust is responsibly maintained.

4. DISCUSSION

As mentioned above, establishing and sustaining trustworthy research requires attention to relationships and systems of accountability. We have seen through these three models of data sharing that approaches to building a chain of trust can take a variety of forms. In the current systems, the bases for trust ranges from hope and established interpersonal relationships to more formalized commonwealths of shared expectations. In the current research environments, each relies to some extent on de-identification or aggregation of patient data as a technical mechanism to advance trust. Essentially the downstream user must be trustworthy enough that stakeholders will release data, and if not *prima facie* trustworthy, then appropriate processes come into play to reduce risk and make the chain of trust more explicit and visible to all parties. DUAs and evolving Honest Broker approaches that depend heavily on forms of de-identification have been doing the bulk of the work in situations where we cannot rely on interpersonal trust networks. As with our regulatory floor, these may provide a start for constructing the chain of trust, but it cannot be sufficient.

The current policy solutions can be characterized as works-in-progress, which has been appropriate given the emergent and dynamic nature of the research and systems in question. However, as the data repositories become more available and in demand, such systems and processes will need to be tested further. As described in this paper, the future of new modes of clinical or genetic data discovery and data sharing needs to deliberately involve policy solutions that sit on top of regulatory approval that in turn leverages technical security, standards and auditability capabilities. We need a community effort to set standards of excellence. There are several proposals for solutions that can leverage the expertise of the computer privacy and security sector, policy makers, bioethicists and clinical informaticians.

These proposals, for the most part, head in two opposite directions: one working to find further, more elaborate ways to

protect data through obfuscation and de-identification on behalf of patients, and another working to personalize and humanize the connections to downstream data users. Depending on the research purpose and scope, either approach could be effective or be complementary to the other; but as we argue at the outset, we need to be cognizant of the scientific and ethical trade-offs of trends toward greater data de-identification as the currently privileged solution. We see the need to separate and advance policies that operationalize the chain of accountability and responsibility through enhanced technically leveraged systems and standards, and practices that change the ways that we treat clinical data in research contexts.

As an alternative to greater de-identification attempts, what would it look like to keep a human face connected to downstream data use? One such approach, tagged as user-centric initiatives, relies on permitting individuals to control their own privacy preferences, data release, and data access (e.g. the company Private Access (www.privateaccess.info) is developing one such approach). A similar project, the UK-based EnCoRe is exploring technical methods to support revocation of consent from patients participating in research [17]. These individualized approaches also have the potential for supporting personalizing reports back to individuals about research that has been conducted with their data or samples, increasing the accountability in the system. We know from emerging social science data that the public is generally divided in their interest in participating at this level in research. 90% of people asked were worried about privacy, but 60% would still participate in a biobank. 48% would give permission for future uses if approved by an oversight board, but an astounding 42% would want to be asked for each use [18]. Individual preferences captured and associated with data as it traverses these systems would permit those who do not care to manage their own data release to give permission for all future uses, while maintaining connection with those who do want more involvement. The same preferences, if persistently maintained, would provide increased quantifiable bases for establishing further technical means to audit provenance and intent, and thus enhance trusted data stewardship.

An additional point of accountability and an opportunity for relationship building is with systems that maintain the ability to return results. This has primarily been discussed as an issue of returning clinically relevant results to individuals; however, further discussion with participants reveals a strong interest in simply knowing where and how their data is being used (a tracking function) and in wanting to hear from researchers high-level reports about how the research is going. Participants want to know their contribution is making a difference, even if it is just to fuel the engine of basic science research. Systems that support researchers capabilities of maintaining greater connection to participants and to track individual data usage would enhance the ability to communicate back to participants what research has been happening with the dataset, including secondary and tertiary uses, as well as potentially providing the ability to enhance the data itself by support patient reported outcomes otherwise inaccessible to the research environment.

Building systems and policy to support these modes will necessarily be iterative and reflective. Feedback loops are a part of regular research practice with longstanding cohort studies (e.g. the Women's Health Initiative or the Framingham Study) where investigators know their investment in relationships with participants is essential to the success of the project over time.

Emerging research paradigms – comparative effectiveness studies, genome wide association studies – will do well to take a lesson from such projects. As discussed, people are willing to take risks and give up some privacy if they know the work that is underway is worthwhile. It is our obligation to help participants see and appreciate the nature of the work and how their data is contributing to research success.

In this paper, we have outlined some of the challenges facing the transition from closed-system data sharing and use environments to more open system environments where end-user relationships cannot be easily standardized in advance. The current dependence on de-identification processes will certainly remain a component of future technically leveraged data sharing processes, at least for quantitative alphanumeric data, but we submit that we need to develop technical-leveraged data management processes that can support persistent information about origin, intent and ownership of clinical data. This capability needs to be meshed with the core floor of regulatory guidance to support new models of governance that can adapt and respond to the unforeseen data management challenges facing biomedicine, and maximize the utility of rich clinically derived data sets for patient participants and researchers alike.

5. ACKNOWLEDGMENTS

This work is supported in part by NIH UL1 RR025014 and DHHS Contract # HHSN268200700031C. Further support for Dr. Edwards was provided by The Greenwall Foundation and the Center for Genomics and Healthcare Equality (NHGRI P50 HG003374).

6. REFERENCES

- [1] Zerhouni EA. Translational research: moving discovery to practice. *Clin Pharmacol Ther.* 2007 Jan;81(1):126-8
- [2] Brickell J, Shmatikov, V. The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. *Knowledge Discovery and Data Mining Conference 2008.*
- [3] Bhumiratana, B, Bishop, M Privacy Aware Data Sharing: Balancing the Usability and Privacy of Datasets, *Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments, June 2009.*
- [4] Nussbaum M. *Poetic Justice: The literary Imagination and Public Life.* Boston: Beacon Press; 1995.
- [5] Milgram S. Behavioral study of obedience. *J of Abnormal Psychology.* 1963;67:371-8.
- [6] Harmon A. Where Did You Go with My DNA. *New York Times.* 2010 April 24.
- [7] Gamble K. High stakes: HITECH's privacy provisions will make costly security breaches even more painful to bear. *Healthc Inform.* 2009;26(7):42-4.
- [8] Yarborough M, Fryer-Edwards K, Geller G, Sharp RR. Transforming the culture of biomedical research from compliance to trustworthiness: insights from nonmedical sectors. *Acad Med.* 2009 Apr;84(4):472-7.
- [9] P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *University of Colorado Law Legal Studies Research Paper*, vol. 09-12, August 13 2009.
- [10] Malin B. A computational model to protect patient data from location-based re-identification. *Artif Intell Med.* 2007 Jul;40(3):223-39.
- [11] Boyd AD, Hosner C, Hunscher DA, Athey BD, Clauw DJ, Green LA. An 'Honest Broker' mechanism to maintain privacy for patient care and academic medical research. *Int J Med Inform.* 2007 May-Jun;76(5-6):407-11.
- [12] Ludman EJ, Fullerton SM, Spangler L, Trinidad SB, Fujii MM, Jarvik GP, et al. Glad You Asked: Participants' Opinions Of Re-Consent for dbGap Data Submission. *J Empir Res Hum Res Ethics.* 2010 Sep;5(3):9-16.
- [13] ONCHIT Meaningful Use Final Rule. 2009. <http://edocket.access.gpo.gov/2010/pdf/2010-17207.pdf>
- [14] Homer N, Szelling S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 2008 Aug;4(8):e1000167.
- [15] Framingham Heart Study [database on the Internet]2010 [cited 10/8/10]. Available from: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v1.p1.
- [16] dbGAP resource [database on the Internet]2010 [cited 10/8/10]. Available from: <http://www.ncbi.nlm.nih.gov/gap/>.
- [17] GTSCB U. EnCore: Ensuring Consent and Revocation. 2010; Available from: <http://www.encore-project.info/>.
- [18] Kaufman DJ, Murphy-Bollinger J, Scott J, Hudson KL. Public opinion about the importance of privacy in biobank research. *Am J Hum Genet.* 2009 Nov;85(5):643-54.