INFORMATICS PROFESSIONALS. LEADING THE WAY.

Prepared Statement by Bradley Malin, Ph.D.
Assistant Professor, Department of Biomedical Informatics, School of Medicine, Vanderbilt University
To the HIT Policy Committee
Privacy & Security Tiger Team
Patient Linking Hearing
Washington, D.C.
Thursday, December 9, 2010

Good morning, and thank you for the opportunity to present testimony this morning. On behalf of AMIA, I am pleased to provide these remarks regarding important issues about patient matching--specifically those most relevant to clinical care and research. We appreciate the opportunity to contribute to these timely policy discussions and wish to emphasize that AMIA very much looks forward to being part of the ongoing efforts as you look to explore, test, implement, and evaluate various possible solutions.

My name is Bradley Malin, and I work as an Assistant Professor of Biomedical Informatics in the School of Medicine and an Assistant Research Professor of Computer Science in the School of Engineering at Vanderbilt University in Nashville, Tennessee. I have experience in developing, applying, and evaluating various record linkage technologies for electronic medical record systems and research datasets. My own research has focused on the construction and evaluation of data privacy models for personal information that is collected, stored, and shared in large complex systems. My remarks are largely based on my personal experiences as a computer science researcher and as a biomedical informatician who has explored these topics for several years.

AMIA is the leading professional association for informatics professionals, serving as the voice of the nation's top biomedical and health informatics professionals and playing an important role in medicine, health care, and science, encouraging the use of data, information and knowledge to improve both human health and delivery of healthcare services.

AMIA is the center of action for more than 4,000 health care professionals, informatics researchers and thought-leaders in biomedicine, health care and science. An unbiased, authoritative source within the informatics community and health care industry, AMIA and its members are transforming healthcare through trusted science, education, and practice in biomedical and health informatics. AMIA is an interdisciplinary and diverse group of individuals and organizations that come from numerous countries, organizations, and backgrounds

working to support and leverage basic and applied informatics principles to help inform public policy issues such as research and evaluation, patient safety, change implementation, and quality of care.

The questions you are considering -- ***What is the problem?*** and ***Why is accuracy in patient matching needed?*** -- are of great interest to me and to AMIA. In the short amount of time available this morning I will highlight several key challenges and implications regarding patient matching. Our submitted written testimony contains additional details and references.

I have organized my remarks to address several overarching questions:

- *In working with large data sets, what are the implications of misidentified data, that is, of being uncertain whether a particular data subject is accurately identified or repeated?*
- *Are particular kinds of research studies hampered more (or less) by the data matching problem?*
- *What percentage of unique individuals in a large data set is incorrectly identified or characterized? How does this percentage vary across differing kinds of databases or datasets?*
- *How does the matching problem differ from other data quality issues?*
- *How do researchers try to correct for matching errors in working with individual level datasets?*

To begin with the first topic:

- ***In working with large data sets, what are the implications of misidentified data, that is, of being uncertain whether a particular data subject is accurately identified or repeated?***

The *impact* of misidentification relies on many factors. Most importantly, it is dependent on the purpose for which the linkage is being performed. There are many different purposes, ranging from patient care to biomedical research to public health efforts. To illustrate the differences, I will take a moment to speak about two environments this morning: patient care and biomedical research.

Before doing so, I will point out there are two distinct types of misidentification, which lead to varying ramifications. The first case of misidentification is the *false match* where it is claimed that two records relate to the same patient, when, in fact, they do not. The second case of misidentification is the *false non-match*, where it claimed that two records fail to relate to the same patient, when, in fact, they do.

More specifically, in the context of clinical care, a false match can cause a patient's record to accumulate erroneous information, such as the medications that the patient has taken or the treatments the patient has received. It is possible that such information could alter the care the patient receives in the present. By contrast, a false non-match results in incomplete knowledge about a patient. This could lead to harm if critical information, such as when medications that could cause adverse interactions are not reported. From an economic perspective, both situations could lead to unnecessary replication of diagnostics or services. Regardless of the impact, it is important to recognize that dirty data in the clinical realm may lead to clinicians' distrust in the historical information that they find in patient records.

Now, I will shift the perspective to the biomedical research environment. In this domain, false matches can affect the system in several ways and depend on the action that results from the match. For example, if the action taken as a result of a record match is to delete one of the records from the system, this could lead to bias in a study. This bias could be positive, such as when the record contained an association with a clinical concept under inquiry. Or this bias could be negative, such as when the record implied there was no association. On the other hand, false matches could lead to improper associations in the system if the records are integrated and not deleted.

In contrast, false non-matches can lead to duplication and fragmentation of information in the research environment. As in the false match case, this could lead to bias in both positive and negative directions. In addition, false non-matches could further lead to fragmentation of a subject's record. The result of fragmentation is that there could be insufficient cases available to study (or detect) certain associations.

- ***Are particular kinds of research studies hampered more (or less) by the data matching problem?***

Any type of research study can be affected by problems in patient matching. As I mentioned earlier, bias can creep into studies as a function of duplication or fragmentation of a subject's records. However, I will point out that longitudinal research studies are of particular concern. In such cases, the fragmentation problem is front and center. For instance, imagine that the subject is diagnosed with a bacterial infection at one institution, but is pronounced as deceased on arrival at a second institution. In this case, we would not recognize that there was a potential association between the bacterial infection and an increased likelihood of mortality.

- ***What percentage of unique individuals in a large data set is incorrectly identified or characterized? How does this percentage vary across differing kinds of databases or datasets?***

Data quality has a direct impact on patient matching accuracy. As a consequence, different types of data sources can vary with respect to the percentage of incorrect identifications. Exact numbers will depend on the individual data set in question. While I do not have sufficient experience to speak about all kinds of databases or datasets (e.g., claims databases, state hospital discharge databases, Medicare, and Medicaid datasets), there is some insight to be gained into misidentification issues through several studies with electronic medical records and epidemiological data.

First, I will mention an investigation at the Regenstrief Institute in Indianapolis, Indiana. The study in question addressed the quality of record linkage using approximately 50,000 newborn screening records and 80,000 medical records in the Indiana Network for Patient Care. Using a state-of-the-art probabilistic matching algorithm, it was observed that over 99% of the records claimed to be a match were, in fact, correct. Additionally, it was observed that approximately 95% of the records that should not have been matched were predicted as non-matches. Dr. Grannis, the lead investigators in this study, is a member of one the panels later today and I believe that he will speak to the Tiger Team about these algorithms in greater detail at that time. [Zhu09]

Second, as a comparison, I will mention a study conducted by researchers at the Washington State Division of Alcohol and Substance Abuse. In this study, they used 600,000 records with approximately 26% duplication. It was observed that off-the-shelf record linkage software was able to achieve 96-97% matching accuracy. [Cambell07]

Third, in an evaluation with medical records from the Vanderbilt University Medical, we illustrated similar statistics. And, though privacy is not the focus of my testimony, I will note that recent research, including our research has suggested that identity obfuscation can be embedded in record linkage schemas to preserve matching accuracy without revealing patient's identifiers. [Durham10, Schnell09]

Again, it should not be assumed that these statistics will hold for all datasets. They are only meant to serve as examples.

- ***How does the matching problem differ from other data quality issues?***

For all intents and purposes, the matching problem is fundamentally a data quality issue. Patient record matching algorithms and software are basically error correction techniques. The

more error or ambiguity that is present in the system, the more difficult it will be derive the true signal (i.e., the correct set of matches).

That said, all data quality issues have some type of semantic bias, and there are certain aspects of the patient matching problem that distinguish it from other data quality issues. First, healthcare is often based on group / family plans. As a result, it is not uncommon for one family member to report his or her own personal information in place of the corresponding information for the actual patient. In research performed at the Regenstrief Institute, it was shown that the Social Security Number, a feature often invoked in record linkage, is sometimes utilized by multiple family members. In particular, it was not unheard of for a parent to report their Social Security Number (SSN) for their child. [Granns02]

Second, record linkage is often dependent on the features that may be misused and overloaded. For example, it has been observed that the same Social Security Number may be shared by many people. There are various reasons why this may occur, but the problem is that misuse and/or sharing of information, such as SSN, is systemic and can have social or geographic biases associated with it.

Third, certain populations are simply more mobile than others. For instance, certain age groups, such as the young adults, are more prone to move from one address to another.

Fourth, there is a problem of "data input for data's sake" in electronic medical record systems. As an example, the registration process for a patient record may require a medical record number and a Social Security Number. However, when the SSN is unknown, some administrators have been known to substitute the medical record number.

As a consequence, the parameters of a matching algorithm may differ from one institution to the next. It may depend on the practices of information collection, validation of the information being used, and simply the distribution of the age groups whose information is being collected and shared.

- ***How do researchers try to correct for matching errors in working with individual level datasets?***

There are several points at which researchers attempt to correct for patient matching errors. First, researchers may try to mitigate the errors themselves, by iteratively refining their matching procedures. In doing so, they will evaluate their matching technique with several datasets, then characterize the bias in their technique, and reevaluate the next generation of the technique with another dataset. Second, when possible, researchers will attempt to use multiple sources to confirm, or triangulate, a match. For instance, imagine record *A* matches to

B and *B* matches to *C*, but the matching algorithm is unsure if A matches to *C*. In this case, it may be inferred by the researcher that *A* and *C* should be labeled as the same patient.

In the event that patient matching errors cannot be resolved through refinement or amendments, correction will depend on the type of matching errors that are expected and the amount of data available. If a sufficiently large amount of data is available, researchers may opt to work with only those records for which they have the highest confidence. For instance, rather than conduct research with all patient records, a researcher may restrict the analysis to only those records for which there is high confidence in its match status. This may reduce the total number of records available to conduct a research study, but it will help to ensure the reliability of the results.

Alternatively, researchers may adjust the threshold upon which they base claims of statistical significance in their scientific discoveries. The amount of adjustment will correlate with the amount of anticipated noise in the matched record sets. From a statistical perspective, this is not an unwise thing to do provided the noise is not very large.

Concluding Remarks and Summary

I would like particularly to thank Shaun Grannis, MD, from the Regenstrief Institute and Elizabeth Ashley Durham from Vanderbilt University for their assistance in preparing this statement. In addition to the citations already noted during these remarks, our written comments include additional relevant references. We would be pleased to help clarify their implications to the points made this morning and to your overall deliberations.

On behalf of AMIA, I would like to thank ONC and the HIT Policy Committee for your attention to an important public policy issue. As a source of informed, unbiased opinions on policy issues relating to the national health information infrastructure, the uses and protection of clinical and personal health information, and a variety of public health considerations, AMIA appreciates the opportunity to contribute to your deliberations. Finally, AMIA again wishes to thank you for convening this meeting and for inviting public comments and testimony. Please feel free to contact us at any time for further clarification of the issues we have raised.

Thank you. I would be pleased to answer any questions that you might have.

For further information, please contact:

Bradley Malin, Ph.D.
Dept. of Biomedical Informatics
Vanderbilt University
2209 Garland Ave, 4th Floor
Eskind Biomedical Library
Nashville, TN 37232

Meryl Bloomrosen, MBA
Vice President Public Policy and Government Relations
AMIA
4519 St. Elmo Avenue Suite 401
Bethesda, Maryland 20814
301-657-5917

Cited References

[Cambell07] K. Cambell, D. Deck, and A. Krupski. Record linkage software in the public domain: a comparison of Link Pus, the Link King, and a "basic" deterministic algorithm. *Health Informatics Journal*. 2007; 14: 5-15.

[Durham10] E. Durham, Y. Xue, M. Kantarciolgu, and B. Malin. Private medical record linkage with approximate matching. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2010: 182-186.

[Grannis02] S. Grannis, M. Overhage, and C. McDonald. Analysis of identifier performance using a deterministic linkage algorithm. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2002: 305-309.

[Schnell09] R. Schnell, T. Bachteler, J. and Reiher. Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*. 2009; 9: 41.

[Zhu09] V. Zh, M. Overhage, J. Egg, S. Downs, and S. Grannis. An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. *Journal of the American Medical Informatics Association*. 2009; 16: 738-745.

Other Selected References

Drozd DR, Lober WB, Kitahata MM, Smith KI, Van Rompaey SE. Developing a relational XML schema for sharing HIV clinical data. AMIA Annu Symp Proc. 2005:943.

Gerdsen F, Müeller S, Jablonski S, Prokosch HU. Standardized exchange of medical data between a research database, an electronic patient record and an electronic health record using CDA/SCIPHOX. AMIA Annu Symp Proc. 2005:963.

Sauleau EA, Paumier JP, Buemi A. Medical record linkage in health information systems by approximate string matching and clustering. BMC Med Inform Decis Mak. 2005 Oct 11;5:32.

Tromp M, Reitsma JB, Ravelli AC, Méray N, Bonsel GJ. Record linkage: making the most out of errors in linking variables. AMIA Annu Symp Proc. 2006:779-83.

Tromp M, Méray N, Ravelli AC, Reitsma JB, Bonsel GJ. Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies. J Am Med Inform Assoc. 2008 Sep-Oct;15(5):654-60. Epub 2008 Jun 25.

Victor TW, Mera RM. Record linkage of health care insurance claims. J Am Med Inform Assoc. 2001 May-Jun;8(3):281-8.