



**Research Report:
Application of IBM Anonymous
Resolution to the Health Care Sector**

February, 2006

*Peter P. Swire
C. William O'Neill Professor of Law
The Ohio State University*

Contents

2	<i>Executive Summary</i>
5	<i>Origin of the Research Project</i>
7	<i>Understanding Entity Resolution and Anonymous Resolution</i>
13	<i>Positive Attributes of Anonymous Resolution to Health Care</i>
13	<i>The Importance of Anonymous Resolution for National Security Sharing</i>
15	<i>The Importance of Privacy and Security to the National Health Information Network</i>
16	<i>Categories of Benefits of Anonymous Resolution</i>
19	<i>Scenarios for Using Anonymous Resolution in Health Care</i>
19	<i>Anonymous Resolution and Linking</i>
22	<i>Anonymous Resolution and Payers</i>
25	<i>Anonymous Resolution and Medical Research</i>
26	<i>Anonymous Resolution and Public Health</i>
28	<i>Conclusion</i>
29	<i>Contacts</i>
30	<i>Appendix: Interviews for this Research Report</i>

Executive Summary

This Research Report examines uses of anonymizing technology for the health care sector. Building on interviews conducted for this paper and the author's experience in privacy, security, and health care, the Report shows how the new technology of Anonymous Resolution (AR, or Anonymous Resolution) shows great promise for solving many pressing health care problems.

The paper first explains the importance in health care of **entity resolution**, or the problem of accurately identifying patients in a world of legacy systems and dirty data. Effective entity resolution will help improve the quality of care, reduce duplicative tests and other unneeded medical procedures, and substantially reduce fraud and other unauthorized payments.

Going beyond traditional entity resolution, **IBM Anonymous Resolution** is a revolutionary technology that enables the transfer of health records or other data (1) across a boundary, such as between two health providers; (2) without the transfer of the name or other personally identifiable information of the individual; and (3) in a mathematical form that permits records of the same individual to be linked together. This paper offers a high-level explanation of how the technology works.

The importance of anonymous resolution has already been accepted for national security applications. It has thus far been less well understood in the health care sector, where the need to share data while protecting privacy is similarly very important. The paper explains the general conditions for where anonymous resolution is likely to help:

1. Sharing of health care fields across a boundary (such as between two or more organizations);
2. Where accurate counting or accurate linking of individual records is useful or where it is useful to be able to re-identify a record after the fact;
3. Where there are legal, business, or other disadvantages to sharing the information in identified form.

Anonymous resolution is the only approach that can do accurate counting and linking without sharing the names and related identifiers of patients. It will prove useful in overcoming many hurdles that might otherwise prevent beneficial information sharing. These hurdles may include: legal barriers to sharing; business reasons why organizations would prefer not to share personally identifiable information (“PII”), such as to protect customer lists; and reducing the burden of computer security and breach notification statutes in systems that hold PII data.

The paper then shows how anonymous resolution can solve key problems in the health care system, in the United States and globally. For example:

- As electronic clinical records are shared at the regional and national level, anonymous resolution is perfectly suited for **linking** patient records held at multiple locations, such as through a Master Patient Index (“MPI”).

- Anonymous resolution has numerous advantages for *payers* in the shift to electronic clinical records. It would improve the quality of care, allow sharing among potential or actual competitors, and facilitate better shopping for health care by employers and plans. It would be especially effective for deterring, detecting, and proving fraud and other unjustified payments. It may also help in the growing move toward “pay for performance.”
- *Medical research* would benefit from anonymous resolution. It would allow researchers to identify records held across organizational boundaries, without identification but with the ability to accurately associate and count the records despite the disparate information silos.
- Anonymous resolution would also be useful for *public health systems*, which can do many of their functions with de-identified data but which in some instances need the ability to associate, count and re-identify data where urgent need exists, such as to track epidemics.

Origin of the Research Project

This Research Report seeks to fill a gap in the current literature. Technologists and policy experts in non-health fields have increasingly understood the importance of anonymous resolution techniques for many kinds of information sharing. In the health care field, however, understanding of and use of anonymous resolution techniques has been relatively limited.

The importance of information sharing technology is especially great in health care at this time. The HIPAA transaction rule pushed the health system in the United States to move payment and other transaction records from paper to electronic form. Only now, however, are we seeing a significant shift of clinical records from paper to electronic form. There are numerous technical, business, and legal challenges as this transition occurs. For the transition to occur smoothly, with public support, it is essential to include effective security and privacy safeguards. Anonymization techniques in many settings appear to be the most cost-effective and privacy-protective ways to facilitate information sharing. Although the emphasis in this paper is often on developments in the United States, the analysis applies similarly to other countries as they expand their use of electronic health records.

The research for this project included detailed interviews with a wide range of health care experts. A list of formal interviews is contained in an appendix. In addition, the author spoke informally to a large number of people on the topic. The people interviewed should not be understood in any way as endorsing this Research Report or any product. Instead, the list is provided to suggest the range of perspectives solicited in the course of the research.

As the author, I agreed to do this project on behalf of IBM Entity Analytic Solutions (EAS) based on belief in the importance of the topic and because I have come to believe that IBM's Anonymous Resolution product is a current example of effective anonymizing technology. Implementation of this product (or other anonymizing technologies if and as they appear) will improve health care while protecting security and privacy. If opportunities are missed to implement anonymizing technologies during the shift to electronic clinical records, then there will likely be weaker security and privacy in the eventual national health information network.

The work for this research project builds on my experience in privacy, security, and health care, both as an academic and government official. In my role as Chief Counselor for Privacy, in the U.S. Office of Management and Budget, I was the White House coordinator for the HIPAA medical privacy rule as proposed in 1999 and issued in 2000. I then returned to law teaching, and am now the C. William O'Neill Professor at the Moritz College of Law of the Ohio State University. I have assisted clients in compliance with HIPAA privacy and security requirements and received grants from the Markle Foundation in connection with three working groups of the Connecting for Health Project. I consulted with Systems Research & Development (SRD) (the original creator of technologies discussed in this paper) on privacy and anonymization before that company was acquired by IBM in 2005 and reconstituted as EAS. My writings and other information are available at www.peterswire.net.

Understanding Entity Resolution and Anonymous Resolution

This section of the Research Report discusses entity resolution before explaining how anonymous resolution works.

Entity Resolution

The problem. A major challenge for the health care system is to identify patients accurately. Accurate “entity resolution” – matching an individual to his or her records – is vital to providing health care and for billing and other purposes. Yet entity resolution is very difficult today, especially when linking the records held by multiple providers, payers, or other organizations.

One source of problems comes in the naming of the individual. There are many “Robert Smiths” in a population. Worse, a particular person might be “Robert,” “Bob,” “Bobby,” and “Robert B.” Smith on different visits or in the records of different providers.

A second source of problems is “dirty” data. Error rates for Social Security numbers apparently often exceed five percent. Addresses and phone numbers often have transposed or otherwise mis-entered data. The numerous paper forms that patients fill out may be illegible or otherwise lead to inaccurate data.

EAS response to the problem. System administrators use many different tools for entity resolution. IBM explains advantages of its entity resolution product compared with other approaches:

“IBM’s DB2 Entity Analytics technology is the industry’s only middleware technology designed to resolve, recognize, disambiguate, and link identities related to a company or organization. Traditional data quality/integration technologies are stove-piped by their inability to integrate information beyond the ‘customer identity.’ They utilize name and address matching, merge/purge, batch processing, and in some cases aggregated master data files to establish identity. Accuracy is impacted by ‘data-drift’ between batch windows, loss of fidelity due to purged attribute data, reduced accuracy resulting from the inability to match on anything beyond two identifiers, and increased privacy concerns due to a reliance on aggregated demographic data sets to establish identity.”

“EAS technology utilizes proprietary resolution process that utilize ‘every’ unique (e.g., SSN) and non-unique (e.g., date of birth or gender) attribute of an individual. The system is refreshed in real time with each new identity element that enters the system eliminating data drift from extended batch windows. EAS maintains the full attribution of the data without ever purging identity attributes so if you have a resolved identity with the same name and address that is really two people, i.e. ‘junior’ and ‘senior’ the system can deconstruct a conjoined identity creating unique identities to correspond to such a revelation with no impact on the depth of originating information because it has always been maintained.”¹

Advantages of effective entity resolution. Health systems, in the United States and globally, are in the midst of an historic shift from paper-based to electronic health records (EHRs). The shift to EHRs means that accurate entity resolution is far more important than ever before. The chief potential benefits of EHRs, notably the ***improved quality of care***, will be eroded or eliminated unless there is accurate linking of records to patients.

¹ The quotation comes from an internal IBM set of FAQs about entity resolution that were provided to the author during the research.

In addition to these patient care and system-wide benefits, effective entity resolution also provides concrete economic benefits to health care providers and payers. Good entity resolution reduces the number of *duplicative tests*, which now are ordered because providers lack the ability to discover existing medical records or lack confidence in whether those tests reflect the true condition of the particular patient. Effective entity resolution will reduce the incidence and cost of *unnecessary medical procedures*. In addition, it will directly reduce *fraud*, such as where patients get prescriptions or other medical care beyond the indicated amount or type.

Anonymous Resolution (AR)

Entity resolution is traditionally done in an identified way. That is, the various organizations who have or share records all know the name of the patient. The focus of this research paper, however, is to understand the potential for “anonymous resolution” (AR) to improve the health care system.

By “anonymous resolution,” the key characteristics are that health records or other data are transferred: (1) across a boundary, such as between two health care providers; (2) without the human readable transfer of the name or other personally identifiable information of the individual; and (3) an irreversible mathematical form that continues to permit records of the same individual to be linked together.

The third characteristic – anonymized linking of individual patient records – is the key innovation of IBM Anonymous Resolution. This paper will next explain technical aspects of how AR works, and then explore applications of AR for the health care sector.

How Anonymous Resolution Works

A more detailed explanation of Anonymous Resolution is contained in the “Inner Workings” section of the May, 2005 whitepaper entitled “IBM DB2 Anonymous Resolution: Knowledge discovery without knowledge disclosure.” This material is available at <ftp://ftp.software.ibm.com/software/data/pubs/papers/db2anonymousres.pdf>.

The key steps in the process are: pre-processing; anonymization through use of one-way hashes; and knowledge discovery through use of the “resolver.”

Pre-processing: Electronic health records or other records are first subjected to pre-processing. Proprietary algorithms prepare the data for anonymization. Notably, ***name standardization*** determines and applies root names (e.g., Rob, Bob, and Bobby equate to Robert). ***Address verification and correction*** compares, verifies, and corrects addresses with U.S. and international address databases. ***Normalization*** applies data-driven rules to addresses, phone numbers, dates of birth, social security numbers, and other significant attributes in preparation for hashing.

Anonymization through one-way hashing. Cryptologists have long used one-way hash techniques to accomplish various security functions, such as the digital signatures used to ensure that a document has not been modified. A one-way hash is basically an algorithm that converts input data into fixed strings of alphanumeric characters. For instance:

<u>Input text</u>	<u>Hashed Value</u>
Dave Travers	h8Z93c7olgwILAAY2uM8

The hash is a “one-way” function because it is mathematically difficult or impossible to figure out the original input (“Dave Travers”) when you know only the hashed value.

At a conceptual level, a one-way hash offers advantages for two organizations that wish to create a more secure environment for data sharing. Organizations A and B can simply one-way hash their data, share the data, and then search for common strings of alphanumeric characters. If the same hash algorithm is used at both sites and the same original text is entered, then this approach would work. In this way, for instance, the two organizations could determine that they shared a particular patient.

In practice, however, this simple approach would rarely succeed. If the original data varies by even one letter, then the hashes become entirely different. Consider the example of three records containing variations, such as Dave Travers, David P. Traverse, and Dave P. Traver:

<u>Input Text</u>	<u>Hashed Value</u>
Record 1	
Dave Travers	h8Z93c7olgwiHCDP2uM8
PPN# 786786543	nZsLGNd3HdsQRpnLONc4
SSN# 027869675	tK8u891GbO6/3DJ1huf6
Record 2	
David P. Traverse	ugis8PSaQkHhCk09IxrU
1 Bourne St	sZw37siaebQ3/jSPXaos
Clinton MA, 01510	hln8OIGbO6/3D76QbFTI
Record 3	
Dave P. Traver	cxke9JSfLoPeRuW4BcmZ
TEL# 5014274475	cdi5Rr1EIDE187KLueVDz
EIN# 896756453	UI7/sdLE87/sFE4G97P

As these examples show, even small variations in first and last name lead to entirely different hashed values. The records would thus appear as three distinct identities. If the objective of sharing the records was to recognize duplicate patients, then matching would not occur and the counting of patients would be inaccurate.

The thus-far unique capability of Anonymous Resolution is its ability to correlate identity data within a hashed data set, despite poor data quality and inconsistencies in how identities are expressed. Working together with the patent-pending pre-processing techniques, described above, the Anonymous Resolution software can recognize ambiguities, misspellings, or partial records within a data set and resolve identities across all attributes. In addition, AR can detect non-obvious relationships between individuals inside of the same anonymized data space.

Knowledge discovery through use of the “Resolver.” Actual matches and other results are produced by the “Resolver,” which can be configured in various ways. The central idea is that the Resolver – the place where matches occur – can receive data from a large number of sources that supply hashed data. The Resolver determines when identities are the same or related and generates messages (e.g., alerts) when necessary, which are passed along to data owners or others who are entitled to receive the results.

The technology is flexible enough to permit differing information sharing models depending on the wishes of participating institutions. For instance, a single organization might use anonymization to facilitate sharing across departments and individuals with different levels of access to sensitive data. Next, in the example above of Organizations A and B, each organization might use the anonymizing technology and have the ability to resolve the common list of patients. Alternatively, a trusted third party might receive the data and operate a Resolver; Organizations A and B would only receive

agreed-upon reports, such as the list of common patients. In this last scenario, Organizations A and B might not receive the anonymized patient records, until and unless there was a match.

Positive Attributes of Anonymous Resolution to Health Care

This section of the Research Report briefly shows how the importance of anonymous resolution has already been recognized in the national and homeland security sectors, providing an important precedent for use in the health care system. It then examines the variety of advantages that anonymous resolution offers.

The Importance of Anonymous Resolution for National Security Sharing

Information sharing in the health care sector can benefit from the experience of information sharing efforts to protect national security in the wake of the attacks of September 11, 2001. There has been widespread and authoritative support for use of anonymizing technologies, such as Anonymized Resolution, in the national security arena. The experience in the national security realm holds important lessons for health care – in both areas, there are compelling benefits from sharing data, but severe privacy and security concerns if the sharing is done in an improper way.

One study on “Anonymization, Data-Matching, and Privacy” was led by Stewart Baker, now Under Secretary for Policy of the U.S. Department of Homeland Security. It stated:

“The thesis of this paper is that cryptography and related technologies will allow democratic nations to make effective use of data-processing capabilities while dramatically reducing the risk of misuse. In particular, advanced techniques for “anonymizing” personal data will help to preserve privacy while obtaining the many benefits of data processing technology.”²

For the U.S. Department of Defense, in the Report of the Technology and Privacy Advisory Committee, a specific recommendation called for a presumption of anonymizing technology:

“Data anonymization—whenever practicable data mining should be performed on databases from which information by which specific individuals can be commonly identified (e.g., name, address, telephone number, SSN, unique title, etc.) has been removed, encrypted, or otherwise obscured.”³

A leader in the debates about information sharing has been the Markle Foundation Task Force on National Security in the Information Age. Its 2003 report highlighted the usefulness of anonymizing techniques for allowing sharing while maintaining privacy:

“[A]nonymizing technologies could be employed to allow analysts to perform link analysis among data sets without disclosing personally identifiable information. By employing techniques such as one-way hashing, masking, and blind matching, analysts can perform their jobs and search for suspicious patterns without the need to gain access to personal data until they make the requisite showing for disclosure.”⁴

Similarly, the promise of anonymous resolution technologies – to facilitate both information sharing and privacy to promote national security – has been emphasized by leading think tanks⁵ and privacy groups.⁶

³ U.S. Department of Defense, The Report of the Technology and Privacy Advisory Committee, “Safeguarding Privacy in the Fight Against Terrorism,” Mar., 2004.

⁴ Markle Foundation Task Force on National Security in the Information Age, “Creating a Trusted Information Network for Homeland Security,” Dec., 2003.

⁵ See Mary DeRosa, “Data Mining and Data Analysis for Counter-Terrorism,” Center for Strategic and International Studies, 2004.

⁶ See James X. Dempsey, Center for Democracy and Technology, Testimony before the House Committee on Government Reform, “Moving from ‘Need to Know’ to ‘Need to Share:’ A Review of the 9-11 Commission’s Recommendations,” Aug. 3, 2004.

The Importance of Privacy and Security to the National Health Information Network

In the health care sector, there has been widespread understanding of the importance of privacy and security protections, especially with the shift to electronic records. To date there has not been, however, the same recognition of the importance of anonymous resolution that we have seen in the national security debates.

In HIPAA, Congress in 1996 recognized that the shift to electronic records for health payments must be accompanied by corresponding privacy and security safeguards. Otherwise, detailed electronic health records would have circulated among providers and health plans without national standards for protecting the data. The HIPAA transaction and code set rule, setting standards for payments and related topics, thus went into effect under the same statutory mandate as the HIPAA privacy and security rules.

The health care sector is now beginning the second major shift to electronic records, this time for clinical records. Privacy and security will once again be important issues. President Bush, in announcing the appointment of David Brailer as National Coordinator for Health Information Technology, said: "This is important for people to understand, that those of us in government who talk about spreading information also, first and foremost, keep your privacy in mind."⁷ The importance of privacy is underscored by opinion polls showing privacy and security as key possible barriers to adoption of a national health information network.⁸

⁷ "President Outlines Health Care Technology Plan for Greater Accessibility," May 27, 2004.

⁸ For instance, the Markle Foundation released a research summary in October, 2005 on "Attitudes of Americans Regarding Personal Health Records and Nationwide Electronic Health Information Exchange." Among the findings was that 91 percent considered the following an absolute or high priority: "The identity of anyone using the system would be carefully confirmed to prevent any unauthorized access or any cases of mistaken identity." See also Alan F. Westin, "How the Public Views Health Privacy: Survey Findings from 1978 to 2005," Feb. 2005.

President Bush has made electronic health records a significant priority. In his State of the Union address in 2006, President Bush said: “We will make wider use of electronic records and other health information technology, to help control costs and reduce dangerous medical errors.” It is thus vital to find effective ways to implement this national priority, including by the use wherever possible of technologies that protect privacy and security.

Categories of Benefits of Anonymous Resolution

The importance of privacy and security to sharing of electronic health records provides the context for how anonymous resolution may solve problems in the health care system. The interviews for this research project underscored the *general conditions for where anonymous resolution is likely to help*:

1. Sharing of health records across a boundary (such as between two or more organizations);
2. Where accurate counting or accurate linking of individual records is useful, or where it is useful to be able to re-identify a record after the fact;
3. Where there are legal, business, or other disadvantages to sharing the information in identified form.

We will briefly examine each of these three conditions, before turning in the next section to specific promising applications in the health care sector.

Sharing records across a boundary. For sensitive data, such as health records, there are numerous legal and other reasons that organizations may be unable or be reluctant to share identified information with other organizations. For instance, under HIPAA and other medical privacy laws, disclosures to third parties are limited without patient authorization. Even in the absence of legal restrictions, organizations may not wish to share identified data for many reasons, including reluctance to share customer lists or the consequences to one's brand equity should there be a large scale unintended disclosure of personally identifiable data.

Accurate counting and linking, or re-identification. There are many times in the health care system that it is important to count patients accurately, such as for outcomes research or epidemiology, to name only two examples. Similarly, there are many times where it is important to link the records of a single patient from multiple sources. Such linking becomes enormously more common with the development of the National Health Information Network.

IBM's ***Anonymous Resolution is the only commercially available approach that can do accurate counting and linking without sharing the names and identifiers of patients.*** Other types of de-identification are similar to AR because they strip out the patient identifiers. However, they are less useful than AR because records of the same patient cannot be linked and each additional record of that single patient appears to be another instance of the disease or other condition. Similarly, such approaches do not make it possible to associate de-identified patient with their prescription or other medical records in outside databases.⁹

⁹ Although the topic is outside of the scope of this paper, my view is that the one-way hashing used in DB2 Anonymous Resolution qualifies as "deidentification" under Section 164.502 of the HIPAA medical privacy rule. Based on my extensive participation in the drafting of that portion of the medical privacy rule, this view is consistent with both the intent and the text of the deidentification provisions.

A related advantage of anonymous resolution is that it permits a patient record to be re-identified under tightly controlled conditions. In situations where there is a showing of need to re-identify a patient record, the organization that originally hashed the identifiers can be asked to do the re-identification. That organization remains in control of the re-identification of the patient.

Legal, business, or other disadvantages of sharing data in identified form – the “pain points.” There are numerous reasons why it may be illegal or unwise to share data in identified form. ***Legal barriers to sharing*** are an obvious first example. HIPAA sets limits on sharing protected health information with other parties. The European Union Data Protection Directive requires member states to comply with comprehensive strict privacy laws, with health records being considered as “sensitive” data subject to stricter standards. Public health agencies and other organizations are often subject to specific confidentiality requirements.

There are many ***business reasons*** why organizations would prefer not to share identified customer data. Even when there are no legal barriers to sharing, businesses may keep records confidential, such as to improve their reputation for trustworthiness or because they think such practices are appropriate. In many settings, businesses may wish to share data for some purposes but do not want to risk their customer lists or other trade secrets.

Increasingly, it is becoming clear that there are ***advantages to holding and transferring records in anonymized ways***. As health records are put into electronic form and placed onto networks, the burden of computer security grows. New security breach statutes

require costly notices to customers when there has been a security breach related to their data. In the face of the growing burdens of holding records in identified form, there are new economic incentives to hold or transfer data in anonymized form where possible.

Put another way, *sharing will often only occur if done in anonymized form*. Much larger and more useful flows of data will often be available if there are assurances that patient identity will not be compromised. The ability to anonymously share clinical data, while maintaining the ability to accurately associate and count, will be a strategic advantage to a growing number of organizations in the increasingly networked health care environment.

Scenarios for Using Anonymous Resolution in Health Care

There are many possible uses of anonymous resolution in health care. Based on the interviews for this project, prominent uses include: for linking health records, such as in a Master Patient Index; as an advantage to payers; for medical research; and for public health. For each category, there is a statement of the problem followed by an explanation of how anonymous resolution can solve that problem.

Anonymous Resolution and Linking

Statement of the problem. In the United States, electronic health records are increasingly being linked at the regional level, often through Regional Health Information Organizations (RHIOs). The next stage will be to link patient records at the national level. In considering the architecture for inter-RHIO linking, there has been longstanding opposition to having a unique health identifier for each American. There is thus enormous pressure to develop alternative architectures that allow linking of patient records held by multiple providers, but without use of a unique identifier. One major pro-

posal is that a Record Locator Service (RLS) should be created with the task of allowing a provider to locate what other providers have records of a specific patient.¹⁰ This RLS would be a form of Master Patient Index (“MPI”), where the records of each patient could be located from multiple sources.

Similar linking challenges exist in other settings. For instance, national health systems in European countries have thus far not typically shared patient records with systems in other countries. As another example, an individual RHIO may decide to use a “federated” model, in which only limited information is held at the center of the RHIO. There may be other settings where two or more organizations wish to link patient records for treatment. For these and other situations, the challenge exists of how to get entity resolution (accurate linking) while preserving privacy and placing limits on the sharing of identified records.

Anonymous identity resolution as a solution. Anonymous identity resolution is perfectly suited for solving these problems. Where records are held by multiple providers, the initial step of entity resolution is very difficult. There are multiple persons with the same name, such as “Robert Smith.” The data is very dirty, with numerous transpositions and other errors, and with widely varying data formats and content. Effective entity resolution, such as that offered by the Entity Analytic Solutions products, is thus a required element of these linking projects.

There are also major advantages to having the entity resolution occur against anonymized data. Having patient identities hashed in the RLS would be a great help in addressing public concerns about the security and privacy of the new record-sharing system. Public

¹⁰ The challenges of linking within the RLS are explored in depth by the Connecting for Health/Markle Foundation report “Linking Health Care Information: Proposed Methods for Improving Care and Protecting Privacy,” (2005), available at http://www.connectingforhealth.org/assets/reports/linking_report_2_2005.pdf. The Anonymous Resolution solution proposed here is consistent with the recommendations of that Markle report.

concerns about a unique health identifier for each American actually led to cancellation of such identifiers in the late 1990's after they had been mandated by HIPAA. To avoid this sort of public backlash in creation of the RLS, it would be simple and effective to state truthfully that the RLS did not even have patient names in its database. Through use of Anonymous Resolution, we could see the creation of the first Anonymous Master Patient Index, or "aMPI."

Holding patient records in the RLS in hashed form would have important computer security advantages. The RLS would then be less of a "honeypot" – an obvious target for computer hackers and identity thieves. If the RLS holds fully identified information, by contrast, then it could become a national registry for American patients, and thus a high-priority target for those wishing to steal detailed personal information. In addition, holding patient records in hashed form would reduce or eliminate the need to send out security breach notices in the event of such a breach – current statutes do not require such notices where the identifying information is encrypted or hashed.

Looked at another way, it is not clear that there are any significant advantages for the RLS to hold the data in identified (unhashed) form. It is not clear that there are any advantages in terms of actual linking that come from the RLS storing the data in unhashed form. In interviews, some persons asked whether the computational expense of hashing and resolving the records would be substantial, slowing system response. Additional interviews with technical experts, however, showed that anonymous resolution is well suited to handle very large record sets. It thus appears that use of anonymous resolution would not noticeably affect the response of the system compared with a non-hashed approach.¹¹

¹¹ Further benchmark testing may be appropriate under realistic conditions in order to verify that there is no noticeable reduction in system response from use of anonymous resolution.

Anonymous Resolution and Payers

Statement of the problem. Payers in the health care system would benefit in numerous ways from greater access to clinical records. As more clinical records shift to electronic form, there is an historic opportunity to make this data available to payers, allowing more reliable analysis and greater accountability. Sharing information among payers, however, is legally often difficult or impossible in identified form. Improvements in data quality and quantity, therefore, will be limited unless there are means to share data in anonymized form that permit accurate association and counting.

A related, and serious, concern is fraud or other unjustified payments in the health care system. Accurate entity resolution is essential to detecting, deterring, and proving fraud. Anonymous resolution is an essential component of fraud reduction in situations where data cannot be shared in identified form.

Anonymous resolution as a solution.

Quality of care and outcomes research. With more data comes more reliable analysis. For the health system as a whole, more data would increase the ***quality of care***, with patients more likely to get the right tests and treatment. One strategy for improving quality of care is ***outcomes research*** – better measurement of the relationships between types of care and outcomes achieved.

The current ability to achieve outcomes research is often limited by the pool of patient records accessible to an individual payer. For instance, a regional HMO might have a million patients enrolled. On such a scale, many potentially important relationships between care and outcome may not be revealed in the absence of larger statistical samples. By contrast, the ability to increase the pool of patient records to ten million or one hundred million persons would allow myriad new non-obvious relationships to be discovered. These rela-

tionships, in turn, feed back into new standards of care, improving quality.

In the absence of anonymous resolution, payers can attempt to do outcomes research with data that is de-identified but lacks the one-way hash. The difficulty with this approach is that there is likely to be a great deal of over-counting. Over-counting occurs, for instance, where a patient with a disease is listed at a home address, with a local provider, and at a tertiary care center. Where this sort of inappropriate counting occurs, the reliability of outcomes research is degraded. Anonymous resolution, with its feature of accurate counting, thus greatly improves outcomes research.

Sharing among potential or actual competitors. HIPAA or other laws place limits on ***sharing among competitors***, unless the sharing is for permitted purposes such as treatment, payment or health care operations. Even where sharing is legally permitted, there may be important business reasons why competitors do not wish to share the names of their customers. In such settings, anonymous resolution may enable worthwhile sharing of medical records while addressing those business concerns.

For example, suppose that two large health plans wished to pool their records in order to do outcomes research or for other data analysis reasons. Business managers at the plans might be reluctant to expose their customer lists to the competitor, for fear that their best customers would be targeted by the competitor. With anonymous resolution, however, the health plans might choose to have a trusted third party merge the records. The records would be transferred to a neutral third party in an anonymized form. This approach would allow accurate entity resolution and then permit data analysis to proceed.

Better shopping by employers, plans, and other payers. Large employers, with their self-insured plans, might use an approach similar to the two health plans just mentioned. Multiple employers in a region, for instance, might pool their data using anonymous resolution. The pooled data could then be used for analysis that would allow the payers to shop more efficiently for health care. By drawing on a larger pool of data, for instance, the payers may learn that some providers cost more per patient than other providers. This data could form the basis for revised ***negotiations on the price of health care, potentially with significant savings.***

Other health plans, which are not employers, could employ similar strategies. These cost savings depend on the ability to pool data that otherwise would not have been pooled, and to get accurate counting even for patients whose care is paid for by more than one payer. In other words, these cost savings precisely match the criteria for anonymized resolution described above.

Reduce fraud and other unjustified payments. A vital cost saving may come from the ***reduction of fraud*** and other unjustified payments. (“Fraud” generally involves a knowing or reckless breach of the rules, whereas other health care services may be improperly charged to a payer without that level of knowledge by the patient.)

Anonymous resolution permits detection of unjustified payments across a wider universe of activity. Where data is pooled together through use of AR, accurate associations and counting make it possible for a payer to detect unauthorized activity that otherwise would not have been visible to the payer. Greater detection and ability to prove fraud, in turn, will deter fraud going forward.

Pay for performance. There is increasing interest in so-called “***pay for performance***” in health care. The idea, supported by many payers, is to shift the focus onto “performance” (health

outcomes) rather than to pay for each procedure done by a provider. For instance, a payer might pay 10% more during a year to a provider who overall performed well compared to less highly-ranked providers.

Accurate measurement of performance will depend on gathering far more clinical data than has historically occurred. Gathering that data in identified form poses large computer security challenges and raises serious privacy concerns. Gathering that data using historical de-identification methods will lead to mis-counting. Anonymous resolution thus offers a unique capability to create accurate pooled data while greatly reducing security and privacy concerns.

Anonymous Resolution and Medical Research

Statement of the problem. HIPAA and other laws can pose a challenge to other medical researchers who wish to review numerous clinical records in the course of their research. Fully identified sharing may require individual authorizations or other large obstacles. Traditional de-identified sharing once again encounters the mis-counting problem.

Anonymous Resolution as a solution.

Records research and clinical trial recruitment. Anonymous Resolution directly responds to these obstacles to effective medical research. In terms of the three criteria for when AR is most effective: (1) researchers need to have records from multiple organizations; (2) with accurate associations and counting; and (3) where sharing in identified form is difficult or impossible. Universities and other research hospitals would benefit greatly from this approach. Epidemiologists and others who use clinical records would expand the range of studies they can conduct. Clinical trials can do recruit-

ment far more efficiently, because they can scan larger populations for appropriate candidates. If and when payers adopt AR, as discussed above, then academic researchers would benefit greatly from having access to the data, while enhancing privacy and security.

Federal statistical agencies, U.K. rules, and other legal barriers to sharing. Interviews uncovered examples of specific legal rules that prohibit some forms of identified sharing. Some U.S. statistical agencies have specific statutes that prohibit such sharing, such as when multiple agencies hold records that may be useful to examine. In such instances, Anonymous Resolution may offer a unique path for conducting research with accurate counting, where counting is obviously of paramount importance for statistical research. Similarly, interviews discovered that the United Kingdom health system apparently has rules calling for de-identification for many “secondary uses” of data, uses outside of the original purpose for which the data were gathered.

These specific laws illustrate a broader point. For research and other purposes, Anonymous Resolution in many settings will be the first commercially viable solution that allows sharing with accurate associations and counting, but without sharing of personally identifiable data.

Anonymous Resolution and Public Health

Statement of the problem. State and national public health systems have compelling reasons to receive various sorts of health data. Health providers are required to report some kinds of incidents, such as gunshot wounds. The increased interest in “biosurveillance” for homeland security purposes has increased the attention given to

innovative ways to offer syndromic surveillance and other reporting to public health agencies. Civil liberties and privacy advocates, along with some providers, are often reluctant, however, to increase greatly the flow of identified health records to such agencies. Beneficial sharing may thus be slowed.

Anonymous Resolution as a solution. Privacy and security concerns are greatest where large flows of PII data are made in identified form. For the vast bulk of public health activities, including syndromic surveillance, there is little or no need on a day-to-day basis for the agencies to know the names of patients. Although HIPAA's Section 512(b) allows disclosure to public health agencies without patient authorization, public health agencies face pressures to preserve confidentiality and reduce the security costs that accompany identified data. Anonymous resolution allows the public health analysis to proceed, and avoids the aforementioned mis-counting problems.

For public health, there are certain circumstances where it is important to be able to re-identify a patient. For example, unusual symptoms may turn out to be linked to a new disease outbreak or an anthrax attack. In such circumstances, it is extremely useful to be able to go back to the entity that originally created the anonymized record to request a re-identification of the identity. Anonymous resolution allows that sort of re-identification, if and only if the entity that holds the original record agrees that re-identification is proper.

Conclusion

As clinical records become available in electronic form, the health care system will face innumerable questions about how to share patients records consistent with security and privacy. In many settings, there are compelling legal, cost, security, and privacy reasons to share records in anonymized form. At the same time, traditional de-identification leads to many errors – one incident of a disease will be over-counted when the patient sees multiple providers, and the ability to link relevant medical records is lost when those records are in more than one database.

Anonymous Resolution is the only commercial product today that can do accurate counting and linking without sharing the names and identifiers of patients. AR thus offers a crucial solution to many pressing problems in the health care system, from a Master Patient Index, to cost control for payers, to improved medical research and public health. All those participating in the next generation of electronic health records should consider how anonymized sharing of patient records can improve health care while preserving privacy and security.

Contact Information

Peter Swire, Author
C. William O'Neill Professor of Law
Moritz College of Law of The Ohio State University
www.peterswire.net

Jeff Jonas, Chief Scientist EAS
(702)851-4697 • jeffjonas@us.ibm.com

John Bliss, Privacy Strategist
(702)851-4683 • jblisslv@us.ibm.com

Beth Drew, Director of EAS Sales
(617)693-1812 • beth_drew@us.ibm.com

Rakesh Goenka, Program Director EAS Marketing
(702)853-4818 • (416)518-2954 • goenka@ca.ibm.com

Additional Information

For the latest information about our products and services, see the following website: www.ibm.com/db2/eas/

Appendix: Interviews for this Research Report

The following persons outside of IBM were interviewed specifically for this project. They are listed for identification only, and their listing here should not be understood as endorsing this Research Report or any product. In addition, the author spoke informally to a large number of persons in connection with this Report.

Casper Bowden, Microsoft

Bill Braithwaite, eHealth Initiative

Claire Broome, Centers for Disease Control and Prevention

Susan Christensen, Agency for Healthcare Research and Quality

Chris Cox, National Center for Health Statistics

John Glaser, Partners Healthcare

Chris Jennings, Jennings Policy Strategies

William Lowrance, consultant on health policy and ethics

Rod Muir, Consultant in Public Health, Scottish National Health Service

Charles Rothwell, National Center for Health Statistics

Clay Shirkey, Markle Foundation

John White, Agency for Healthcare Research and Quality

The following persons inside of IBM participated in interviews and related research for this report:

Craig Bennett

John Bliss

Nikole Burroughs

Dawn Campbell

Dina Canarozzi

Boaz Carmeli

Simona Cohen

Adam Crafton

Doc Dockter

Kathy Dodsworth-Rugani

George Eisenberger

Nancy Friedland

Monty Green

Paul Grundy

David Hom

Joseph Jasinski

Jeff Jonas

Kevin Julier

Shawna Koch

James Kaufman

Emely Marcano

Dana Murphy

Ivo Nelson

Betty Norton

Karen Phelps

Beverly Phillips

kathleen Pizzarello

Amanda Spaulding

Rick Stevens

Mike Svinte

Evelyn Torzilli

Karen Witting

Susan Zimmermann



© Copyright IBM Corporation 2006


IBM (United States of America)
Entity Analytic Solutions
6600 Bermuda Rd, Suite A
Las Vegas, Nevada
United States of America, 89119

Printed in the United States of America
02-06
All Rights Reserved.

DB2, IBM, the IBM logo, and the On Demand logo
are trademarks of International Business Machines
Corporation in the United States, other countries or both.

Other company, product and service names may be
trademarks or service marks of others.

References in this publication to IBM products or
services do not imply that IBM intends to make them
available in all countries in which IBM operates.

 Printed in the United States of America on recycled paper
containing 10% recovered post-consumer fiber.

IBM's customers are responsible for ensuring their own compliance with relevant laws and regulations. It is a customer's sole responsibility to obtain advice of competent legal counsel as to the identification and interpretation of laws and regulations that may affect a customer's business and any actions required to comply with such laws. IBM does not provide legal, accounting or audit advice or represent or warrant that its services or products will ensure that a customer is in compliance with any law.