Privacy in Context: The Costs and Benefits of a New Deidentification Method

by

Stanley Trepetin

B.A., Computer Science and Mathematics
Cornell University, 1989

M.A., Liberal Studies
Duke University, 1999

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN HEALTH INFORMATICS
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2006

Signature of Author:____________________________________
Department of Electrical Engineering and Computer Science
May 11, 2006

Certified by:____________________________________
Peter Szolovits
Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by:____________________________________
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Privacy in Context: The Costs and Benefits of a New Deidentification Method

by

Stanley Trepetin

Submitted to the Department of Electrical Engineering and Computer Science on May 11, 2006 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Health Informatics

Abstract

The American public continues to be concerned about medical privacy. Policy research continues to show people's demand for health organizations to protect patient-specific data. Health organizations need personally identifiable data for unhampered decision making; however, identifiable data are often the basis of information abuse if such data are improperly disclosed. This thesis shows that health organizations may use deidentified data for key routine organizational operations.

I construct a technology adoption model and investigate if a for-profit health insurer could use deidentified data for key internal software quality management applications. If privacy-related data are analyzed without rigor, little support is found to incorporate more privacy protections into such applications. Legal and financial motivations appear lacking. Adding privacy safeguards to such software programs apparently doesn't improve policy-holder care quality. Existing technical approaches do not readily allow for data deidentification while permitting key computations within the applications.

A closer analysis of data reaches different conclusions. I describe the bills that are currently passing through Congress to mitigate abuses of identifiable data that exist within organizations. I create a cost and medical benefits model demonstrating the financial losses to the insurer and medical losses to its policy-holders due to less privacy protection within the routine software applications. One of the model components describes the Predictive Modeling application (PMA), used to identify an insurer's chronically-ill policy-holders. Disease management programs can enhance the care and reduce the costs of such individuals because improving such people's health can reduce costs to the paying organization. The model quantifies the decrease in care and rise in the insurer's claim costs as the PMA must work with suboptimal data due to policy-holders' privacy concerns regarding the routine software applications. I create a model for selecting variables to improve data linkage in software applications in general. An encryption-based approach, which allows for the secure linkage of records despite errors in linkage variables, is subsequently constructed. I test this approach as part of a general data deidentification method on an actual PMA used by health insurers. The PMA's performance is found to be the same as if executing on identifiable data.

Acknowledgements

A variety of people should be thanked for the completion of this work. I must thank MIT Professor Peter Szolovits for his wisdom, endless patience, and the freedom which he provided me as I explored ideas that ultimately made sense to me. I owe him significant debt. I must thank Meghan Dierks, on faculty at Harvard Medical School, for helping me with the clinical and insurance aspects of this work. Her assistance got me much closer to understanding the financial and operational aspects of US health insurance organizations, grounding my work in a realistic US healthcare perspective. I really thank you for the advice, Meghan. I must also thank MIT Professors Nazli Choucri and Hal Abelson for their social science and mathematical suggestions, respectively. Both Professors gave me advice on the cohesion and rigor of this work, and how to make it both human and analytical simultaneously. I am very thankful for the help, Professors. Different individuals who helped me along the way, either by reviewing parts of my work or being a sounding board as I developed my thoughts, must also be thanked. They include MIT Professors Ronald Rivest and Shafi Goldwasser; MIT graduate students Matthew Lepinski, Nick Harvey, and David Woodruff; University of Louisville Professor Marianne Hutti; and all the past and current students within the Clinical Decision-Making Group at the Computer Science and Artificial Intelligence Laboratory at MIT of which I am part. These students include: Mike McGeachie, Ronilda Lacson, Delin Shen, Lik Mui, Min Wu and a number of others who at times only stopped by for a brief conversation but whose suggestions were still useful for my work. I appreciate the advice, friends! I am also very thankful for Susan Spilecki's editing advice regarding the thesis.

I must also thank those who may not have contributed to this work directly but still wished me well when our paths crossed in different contexts during my MIT career. My friends who I knew for many years, including my best friend Michael Koss, or others who I met over the past 6 years, thank you for the encouragement. There are too many of you to mention but I can say that your many "Good Luck" wishes have had the intended impact.

I must also thank all those whom I've interviewed for this thesis. Thank you for your time; your assistance is appreciated.

Finally, I must thank my parents, especially my Dad, without whose support the completion of this long, extremely challenging, and personally-changing endeavor would have been much more difficult. Thank you, Mom and Dad for your continued support—love you always!

This work was supported by a National Library of Medicine (NLM) Biomedical Informatics Training Grant (LM 07092) and an MIT Research Assistantship funded by the National Multiprotocol Ensemble for Self-scaling Systems for Health (N01-LM-3-3515), also a program of the NLM.

1 Introduction

1.1 Health Privacy Concerns Continue

Individuals continue to be concerned about medical privacy.[1] [2] [3] [4] [5] As a number of commentators indicate, privacy, in general, refers to information control.[6] [7] [8] [9] [10] A consumer should be able to control information available about her. Concerns have festered for over a decade about how health organizations inadvertently publicize sensitive information, improperly dispose of protected health information (PHI), or improperly use software to manage PHI, undermining data control.[11]

[1] Center for Democracy and Technology, "Statement of Janlori Goldman, House Committee on Government Reform and Oversight," 1996, <http://www.cdt.org/testimony/960614goldman.html> (9 October 2003).

[2] Health Privacy Project, "Health Privacy Polling Data," 2001, <http://www.healthprivacy.org/content2310/content.htm> (9 October 2003).

[3] Janlori Goldman and Zoe Hudson, "Virtually Exposed: Privacy and E-health," *Health Affairs*, 19 (2000): 141.

[4] Harris Interactive, "Privacy On and Off the Internet: What Consumers Want," 2002, 64-65, <http://www.aicpa.org/download/webtrust/priv_rpt_21mar02.pdf> (10 October 2003).

[5] HIPAAps Privacy and Security, "Examples of Privacy Violations," 2003, <http://www.hipaaps.com/main/examples.html> (31 March 2005).

[6] See Donna L. Hoffman, "The Consumer Experience: A Research Agenda Going Forward," 14 May 2003, <http://elab.vanderbilt.edu/research/papers/pdf/manuscripts/FTC.privacy.pdf> (31 March 2005).

[7] Tamara Dinev, "Privacy Concerns and Internet Use – A Model of Tradeoff Factors," <http://wise.fau.edu/~tdinev/publications/privacy.pdf> (31 March 2005).

[8] Eve M. Caudill and Patrick Murphy, "Consumer Online Privacy: Legal and Ethical Issues," *Journal of Public Policy & Marketing*, 19 (2000): 10.

[9] Secretariat, Treasury Board of Canada, "So, What Exactly Is Privacy?" 26 September 2003, <http://www.cio-dpi.gc.ca/pgol-pged/piatp-pfefvp/course1/mod1/mod1-2_e.asp> (31 March 2005).

[10] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke, "Information Privacy: Measuring Individuals' Concerns About Organizational Practices," *MIS Quarterly*, 20 (1996): 172, 181.

[11] In this thesis, we follow the definition of the Health Insurance Portability and Accountability Act (HIPAA) in defining "protected health information" (PHI): individually identifiable health information relating to a physical or mental health condition of an individual, the provision of his care, or the payment for that care, as will be discussed later on in the text. (Taken from Centers for Disease Control and Prevention, "HIPAA Privacy Rule and Public Health," 2003, <http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm> (3 April 2005)). Not all organizations define PHI with similar specificity or content. (Office of Technology Assessment, *Protecting Privacy in Computerized Medical Information* (Washington, DC: US Government Printing Office, 1993), 2-5; American Association of Health Plans, "Statement on the Confidentiality of Medical Information and the Medical Information Protection Act of 1998," 1998, <http://www.aahp.org/Content/ContentGroups/Testimony/Confidentiality_and_Protection_of_Medical_Information_(Feb__26,_1998).htm> (31 March 2005); American Medical Association, "The Ethical Force Program," December 2000, 13, <http://www.ama-assn.org/ama/upload/mm/369/ef_privacy_rpt.pdf> (2 September 2005)). Nevertheless, all health organizations recognize the need to protect health information in a secure manner. In this thesis, "PHI" will be used to aggregate related terms.

1.1.1 Legal and Organizational Privacy Protections

Until recently, the laws requiring organizations to protect PHI have been inconsistent. Federal and state regulations were fragmented, addressing different entities that manage information, individuals, practices, or medical conditions without broad comprehensive solutions.[12] [13] [14] For example, the 1997 Balanced Budged Act required *Medicare+Choice* organizations to create safeguards to protect personally identifiable information.[15] The *Veteran's Benefits Section* of US law provides for medical record confidentiality when involving cases of *drug abuse, HIV infection, or sickle cell anemia.*[16] In February 2000, President Clinton's Executive Order banned usage of *genetic information* in *federal hiring and promotion* decisions.[17] Almost all states have specific laws to protect *genetic information* and certain health conditions including *mental illness, communicable diseases, and HIV/AIDS.*[18] [19] However, only about *half* of states have a general law that prohibits one entity from disclosing health information to another without patient authorization.[20]

Legal consensus has appeared in the form of the Health Insurance Portability and Accountability Act (HIPAA) regulation which went into force on April 14, 2003. HIPAA offers basic national health privacy protection to individuals across several types of health organizations as opposed to the prior inconsistent protection in the US.[21] We will discuss HIPAA and its effectiveness later on in the text.

Health organizations try to protect privacy. The American Association of Health Plans (AAHP) represents health plans—organizations that pay for and manage care—covering 170 million lives.[22] In 1998 comments on federal privacy legislation, AAHP stated that health plans already follow AAHP's "Code of Conduct."[23] The Code follows federal and state laws, and AAHP member plans are required to provide safeguards for PHI, confidentiality training to staff, and a disciplinary policy for employee non-compliance.

[12] US Department of Health and Human Services (HHS), "Standards for Privacy of Individually Identifiable Health Information" (part 1), 28 December 2000, 82469, 82473, <http://www.hhs.gov/ocr/part1.pdf> (10 October 2003).
[13] Health Privacy Project, "Exposed: A Health Privacy Primer for Consumers," 1999, 5, <http://www.healthprivacy.org/usr_doc/34775.pdf> (13 October 2003).
[14] See Health Privacy Project, *State Privacy Law Summaries*, <http://www.healthprivacy.org/info-url_nocat2304/info-url_nocat_search.htm> (13 October 2003).
[15] HHS (part 1), 82469.
[16] HHS (part 1), 82469.
[17] HHS (part 1), 82469.
[18] Health Privacy Project, "Exposed: A Health Privacy Primer for Consumers," 5.
[19] See also different state laws in Health Privacy Project, *State Privacy Law Summaries*.
[20] HHS (part 1), 82473.
[21] HHS (part 1), 82463-82464.
[22] American Association of Health Plans, "About AAHP," 2003, <http://www.aahp.org/template.cfm?section=About_AAHP> (12 October 2003).
[23] American Association of Health Plans, "Statement on the Confidentiality of Medical Information and the Medical Information Protection Act of 1998."

Many health plans have already implemented PHI security safeguards such as unique log-on passwords, audit trails, and PHI access based on a "need to know" rationale.[24] [25] As another example, the National Association of Insurance Commissioners (NAIC) aims to protect insurance consumers but also maintain a competitive insurance industry in each state.[26] The NAIC adopted a Health Information Privacy Model Act in 1998. A model for state legislation, it allows for individuals to examine PHI and modify it under certain circumstances, and requires insurers to provide security for health information and send a notice to individuals about how health information will be used.[27]

1.1.2 Privacy Protection Connected to Finance and Healthcare

Despite laws and organizational practices violations persist. The following types of problems were reported by local newspapers from around the country over the past decade:[28]

1) Accidents, such as when about 400 pages of detailed psychological records of at least 62 youths were accidentally posted on the University of Montana Web site in 2001;
2) Unclear privacy policies, such as when a patient at Brigham and Women's Hospital in Boston, MA found that employees had accessed her medical record more than 200 times in 2000;
3) Poor security practices, such as when a Tampa, FL public health worker took a computer disk containing the names of 4,000 HIV positive people from work and sent the disk to two newspapers in 1996.
4) Unclear data sharing practices, such as when in 1995 in Pennsylvania, the drugstore Rite-Aid provided to a state agency information about prescriptions being filled by agency employees, including one employee's medication indicative of his HIV positive status. In the employee's unsuccessful ensuing lawsuit against the agency, the court indicated that the employer should be allowed to know details of how its employees used the health plan.

Health organizations may be struggling to maintain profitability and not impact quality of care while protecting privacy. Incorporating technologies and policies which protect privacy appears to undermine net income. In a 1997 study, several health organizations indicated that the protection of health information does not serve as a market differentiator.[29] Others believe the financial costs of installing privacy and security

[24] HHS (part 1), 82478.

[25] American Association of Health Plans, "Statement on the Confidentiality of Medical Information and the Medical Information Protection Act of 1998."

[26] National Association of Insurance Commissioners, "NAIC Mission Statement," <http://www.naic.org/about/mission.htm> (31 March 2005).

[27] American Medical Association, "The Ethical Force Program," 11.

[28] HIPAAps Privacy and Security.

[29] National Research Council, *For the Record: Protecting Electronic Health Information* (Washington, DC: National Academy Press, 1997), 156.

safeguards are too "burdensome."[30] Quality of care might suffer as proper diagnosis or treatment decisions are undermined due to limited information flow.[31]

1.2 Outline of Thesis

This thesis explores such questions. Does incorporating privacy protections restrict profitability or impair improvement of care? This thesis will show that, if data are not analyzed with rigor, it is difficult to show financial and quality of care benefits associated with installing privacy-protecting policies and technologies, and this difficulty undermines the adoption of such safeguards. However, if data are analyzed in more depth, financial and care improvement benefits may be shown, which should encourage the use of privacy-protecting practices.

In this thesis, I will create a technology adoption and decision analytic model for a hypothetical but realistic for-profit health insurer. The term *insurer* will refer to health insurer throughout this thesis. I will examine a subset of the insurer's internal software applications used for key routine operations. There is no legal requirement to provide extended privacy protection within these software applications, which help reduce the insurer's expenses and enhance the care of its policy-holders. To examine the incentives for the adoption of privacy-protecting practices, I will examine the impact of adding a privacy-enhancing technology, which I design in this thesis, on the insurer's profitability and ability to support the improved care of its policy-holders. First, I will show how an unrigorous data analysis demonstrates to the insurer no financial or health improvement benefits of adding extra privacy protection to the routine applications. Second, I will explore data in more depth to show how financial and quality of care benefits may exist. Third, I will create a threat framework describing the protections that must be offered to data depending on their organizational context. If data are to be protected by the same organization that generated them, as is the case for our routine software applications, more data protections must be created. Organizational employees may know some of the security processes involved in the data protection, against which a security approach must guard. Fourth, I will describe my new privacy-enhancing technology, which preserves privacy during data linkage in software applications in general while handling errors of organizationally- internally- and externally-generated data. Finally, I will test this technology by incorporating it into one of the applications used by the insurer, one of the key routine software applications we examine, and investigate the application's performance on its required computations.

[30] Janlori Goldman, "Protecting Privacy to Improve Health Care," *Health Affairs*, 17 (1998): 51.
[31] US Department of Health and Human Services (HHS), "Standards for Privacy of Individually Identifiable Health Information" (update), 14 August 2002, 53209, <http://www.hhs.gov/ocr/hipaa/privrulepd.pdf> (24 September 2005).

1.3 Privacy Concerns within Health Organizations

The analysis below demonstrates how an unrigorous investigation finds no benefits to health organizations from adding more privacy protections. We describe a context in which a set of health organizations currently do not protect privacy as much as may be possible. We explore when privacy-protecting practices promoted by standards organizations are not fully implemented by health organizations. Health privacy is instantiated within organizations by a set of managerial and technical processes to ensure PHI protection. We will rely on the practices suggested by the Ethical Force Program (EFP). The EFP is a program of the American Medical Association and aims to improve health care by encouraging more ethical behavior among all stakeholders.[32] The EFP tenets, below, are useful as they provide for considerable consumer control over PHI, which forms the basis of individuals' understanding of medical privacy, as described in Section 1.1. EFP's tenets were also adapted from the more widely accepted Fair Information Practices.[33] Many national and international organizations have adopted versions of the Fair Information Practices for their own constituents concerning data management and privacy.[34] Based on a 1998 expert panel convened by the EFP, comprised of a variety of physician, hospital, patient, insurance, legal, public health, and IT leaders and experts, in 2000, the EFP issued the following set of recommendations detailing how to protect PHI:[35] [36] [37]

1) Trustees. The following practices apply to all "health information trustees." A trustee is an individual or organization that creates, stores, transmits, or uses PHI.[38]
2) Deidentification of PHI. If the PHI is adequately deidentified, then the health information trustees may use the data; there should be no associated privacy concerns.
3) Transparency. Health information trustees should make publicly available explanations of their policies and procedures for the collection and use of PHI.
4) Consent. Whenever possible, health information trustees should obtain informed consent from individuals with regard to the collection, use, and storage of their PHI. Otherwise a publicly formal process should be used to waive such consent.
5) Collection Limitation. Health information trustees should limit the collection of health information only to current needs or reasonably anticipated future needs that are made explicit at the time of consent.
6) Security. Health information trustees should protect PHI using reasonable means of security. An internal security program guiding such decisions should be established.

[32] American Medical Association, "About the Ethical Force Program," 18 July 2005, <http://www.ama-assn.org/ama/pub/category/14401.html> (17 September 2005).
[33] American Medical Association, "The Ethical Force Program," 10.
[34] American Medical Association, "The Ethical Force Program," 10-12.
[35] American Medical Association, "The Ethical Force Program," 3-5.
[36] Matthew K. Wynia, Steven S. Coughlin, Sheri Alpert, Deborah S. Cummins, Linda L. Emanuel, "Shared Expectations for Protection of Identifiable Health Care Information," *Journal of General Internal Medicine*, 16 (2001): 100.
[37] American Medical Association, "The Ethical Force Program," 6.
[38] American Medical Association, "The Ethical Force Program," 13.

7) Individual Access. People should be granted viewing and copying rights concerning their PHI. They may amend their PHI if the information appears incorrect.[39]
8) Data Quality. Health information trustees should seek to ensure that the PHI in their care is accurate and up-to-date, including conducting periodic data accuracy audits.[40]
9) Information Use Limitation. Health information trustees should limit the disclosure and use of PHI to purposes made explicit at time of consent or by authorization via a publicly accountable formal process, such as in step 4 above.
10) Accountability. Policies should exist to ensure that health information trustees be accountable for adhering to the standards for the collection, storage, and use of PHI, including the responsible transfer of such data to other accountable information trustees.

1.3.1 Basic Quality Management Software Applications

We find that a set of internal software applications within for-profit insurers might not fully follow such practices. A health insurer can be defined as an organization that pays for the care offered to patients by providers—physicians, clinics, hospitals, or pharmacies—often by processing health care claims.[41] [42] Often a health insurer "manages" the care received by patients such as by entering into various contractual agreements with provider organizations to share the "risk" associated with patient care.[43] [44] We will examine what I call the insurer's basic quality management applications (BQMA) which the insurer uses to monitor its organizational efficiency. These are internal software applications that link and compute results based on medical and pharmacy claims data.[45] [46] [47] There may be several BQMA within an average insurer. We look at four common applications:

[39] American Medical Association, "The Ethical Force Program," 19.

[40] American Medical Association, "The Ethical Force Program," 20.

[41] See National Research Council, 66.

[42] See The Kansas Department of Health and Environment, "Charitable Health Program Overview," <http://www.kdhe.state.ks.us/olrh/CHPoverview.htm> (31 March 2005).

[43] National Research Council, 66.

[44] W.K. Kellogg Foundation, "Frequently Asked Questions...Insurance and Managed Care," 2, <http://www.wkkf.org/Pubs/Devolution/NCSL_FA_Insurance_and_managed_care_00331_02768.pdf> (2 September 2005).

[45] See American Health Information Management Association, "Sizing up HEDIS: Experts Take System's Measure," 2002, <http://library.ahima.org/xpedio/groups/public/documents/ahima/pub_bok1_009714.html> (12 October 2003).

[46] FACTS Services, Inc, "Products At-a-glance," <http://factsservices.com/products/products_glance.asp> (31 March 2005).

[47] See DxCG, "Disease Management and Quality Improvement Report," May 2003, <http://www.dxcg.com/press/DMQualityReport.pdf> (31 March 2005). For example, the Ingenix Procise Predict product mentioned in this DxCG article is a Predictive Modeling software platform which analyzes claims and is one of the BQMA based on how I define the BQMA, in the text. (See Ingenix Corporation,

1) Utilization Review. This application permits the insurer to ensure that relevant and cost-appropriate medical care is given by a provider to a policy-holder. For example, an important purpose of managed care organizations is to control growing US health care expenditures.[48] Insurance organization analysts review policy-holder care and costs to authorize payment only for those treatments that meet appropriate cost-management and treatment guidelines.[49]
2) Provider Profiling. This is similar to Utilization Review but focuses on the providers. The insurer's staff examines information regarding provider practices. The intent is to comprehend and potentially influence providers connected with the insurer.[50] Provider Profiling information may be used to select which doctors become part of the insurer's managed care network, or possibly even investigate inappropriate treatment prescribed by providers.[51]
3) Health Plan Employer Data and Information Set (HEDIS). HEDIS is a collection of various organizational operational data used to quantify the insurer's performance. Statistics on breast cancer screening, births, customer satisfaction, and other measures are captured by HEDIS.[52] Such measures may be publicized so that individuals and organizations can compare insurer performance such as when purchasing health benefits.[53] [54]
4) Predictive Modeling (PM). PM attempts to identify people for disease management. Disease management is a set of clinical and management protocols to improve the health status and reduce costs of individuals with chronic conditions such as congestive heart failure or chronic obstructive pulmonary disease.[55] PM examines claims data to predict which individuals, without intervention, might have worsening health status and costs at a future period.[56] [57]

"Improve Your Medical Management and Underwriting Effectiveness," <http://www.ingenix.com/esg/products.php?pid=10> (20 April 2005)).

[48] Alain C. Enthoven and Sara Singer, "The Managed Care Backlash and the Task Force in California," *Health Affairs*, 17 (1998): 95-6.

[49] See Thomas G. Kremer and Ellis Gesten, "Confidentiality Limits of Managed Care and Clients' Willingness to Self-Disclose," *Professional Psychology: Research and Practice*, 29 (1998): 558.

[50] Bettermanagement.Com, "Effective Provider Profiling: Enhancing Care, Improving Costs," Webcast reviewed on December 16, 2004.

[51] Bettermanagement.Com.

[52] National Committee for Quality Assurance, "HEDIS 2005 Summary Table of Measures and Product Lines," <http://www.ncqa.org/Programs/HEDIS/HEDIS%202005%20Summary.pdf> (26 March 2005).

[53] Implied, National Committee for Quality Assurance, "The Health Plan Employer Data and Information Set (HEDIS)," <http://www.ncqa.org/Programs/HEDIS/> (12 October 2003).

[54] Joseph W. Thompson, Sathiska D. Pinidiya, Kevin W. Ryan, Elizabeth D. McKinley, Shannon Alston, James E. Bost, Jessica Briefer French, and Pippa Simpson, "Health Plan Quality-of-Care Information Is Undermined by Voluntary Reporting," *American Journal of Preventive Medicine*, 24 (2003): 69.

[55] Disease Management Association of America, "Definition of Disease Management," 2003, <http://www.dmaa.org/definition.html> (14 October 2003).

[56] See Case Western Reserve University, "Disease Management Programs," <http://www.case.edu/med/epidbio/mphp439/Disease_Management.htm> (2 September 2005).

[57] Privacy Sector Advocacy, "Disease Management and Chronic Diseases," November 2002, 2.

[58] Disease management staff intervenes with the identified individuals using standards-based care to prevent health decline.

1.3.1.1 Disease Management and Predictive Modeling

We will explore privacy concerns regarding disease management and PM, extending our analysis to the other BQMA. Note, one assumption we make in this thesis is that PM software is used to identify disease management candidates. There are other approaches to identifying such candidates, including physician referral and a review of the patient's medical record.[59] [60] However, using software to scan claims may be more tenable for an insurer. For example, in some cases, the health plan is organizationally separate from the physicians providing care.[61] Thus having access to medical records or obtaining physician referral might be less possible for the insurer.

We describe disease management and PM in depth. The Disease Management Association of America defines disease management as "a system of coordinated healthcare interventions and communications for populations with conditions in which patient self-care efforts are significant."[62] Disease management arose from the need to manage chronic patient care and associated costs better than they were being managed. Today, caring for the chronically ill consumes a disproportionate share of national health spending.[63] By 2010, an estimated 120 million Americans will have chronic conditions.[64] In the past ten years, disease management programs have arisen to ensure that health professionals and patients alike follow "evidence-based guidelines" and that patients are encouraged to monitor their own care.[65] A key reason for current suboptimal care is inconsistent care. Evidence-based guidelines exist for many of the common chronic conditions such as diabetes or acute lower-back pain. However, the US health care system is decentralized and such standards are not always followed.[66] Patients themselves

[58] Susan L. Norris, Phyllis J. Nichols, Carl J. Caspersen, Russell E. Glasgow, Michael M. Engelgau, Leonard Jack, Jr, George Isham, Susan R. Snyder, Vilma G. Carande-Kulis, Sanford Garfield, Peter Briss, and David McCulloch, "The Effectiveness of Disease and Case Management for People with Diabetes. A Systematic Review," *American Journal of Preventative Medicine*, 22 (2002), 20.

[59] W. Pete Welch, Christopher Bergsten, Charles Cutler, Carmella Bocchino, and Richard I. Smith, "Disease Management Practices of Health Plans," *The American Journal of Managed Care*, 8 (2002): 358.

[60] Joshua J. Ofman, Seonyoung Ryu, Jeff Borenstein, Stephen Kania, Jay Lee, Amy Grogg, Christina Farup, and Scott Weingarten, "Identifying Patients with Gastroesophageal Reflux Disease in a Managed Care Organization," *American Journal of Health-system Pharmacy*, 58 (2001): 1608.

[61] Pacific Business Group on Health, "Disease Management Effectiveness Project," November 2002, 4, <http://www.pbgh.org/programs/dmep/disease_mgmt_report_11-02.pdf> (2 September 2005).

[62] Disease Management Association of America.

[63] California Healthcare Foundation, "E-disease Management," November 2001, 6, <http://www.chcf.org/documents/ihealth/EDiseaseManagement.pdf> (2 September 2005).

[64] Robert Wood Johnson Foundation, "Chronic Care in America: A 21st Century Challenge," November 1996, <http://www.rwjf.org/files/publications/other/ChronicCareinAmerica.pdf> (24 February 2006).

[65] Disease Management Association of America.

[66] Institute of Medicine, *Crossing the Quality Chasm* (Washington, DC: National Academy Press, 2001), 28.

don't always comply with necessary treatments.[67] By following standards-based care, a disease management program hopes to improve a patient's clinical status.[68] Further, costs should be reduced as maintaining the health of healthier people may cost less.[69] [70] Some clinical and economic successes have been reported concerning certain chronic conditions, such as diabetes and asthma.[71] Still, the overall business case has not been developed for all health contexts.[72] [73] [74]

Claims data are duplicated for PM. After the claim is paid by the insurer, it is stored in a master database or file. Ultimately, a copy of this record is made available to PM.[75] In this thesis we will call such a secondary database the copy data store. PM is run against the copy data store to identify disease management candidates.

After identification, PM frequently stratifies the individuals based on their "risk."[76] The purpose is to align limited resources commensurately with the "risk" presented by such patients; the insurer can optimize its resources.[77] [78]

1.3.2 Basic Quality Management Applications Privacy Concerns

PM uses identifiable data and some people have concerns about the security of such data within organizations.[79] Identifiable data are needed for PM operations. Data values must

[67] Government of British Columbia, "Chronic Disease and Your Health: Information for Patients," *Chronic Disease Management*, 2003, <http://www.healthservices.gov.bc.ca/cdm/patients/index.html> (10 October 2003).

[68] Case Western Reserve University.

[69] Case Western Reserve University.

[70] Privacy Sector Advocacy, 2.

[71] American Association of Health Plans/Health Insurance Association of America, "The Cost Savings of Disease Management Programs: Report on a Study of Health Plans," November 2003, <http://www.aahp.org/Content/ContentGroups/Homepage_News/Disease Management_Short_Report.doc> (25 August 2005).

[72] Norris, 20.

[73] Pacific Business Group on Health, 4.

[74] Geoffrey B. Baker, "Integrating Technology and Disease Management : The Challenges," *Healthplan*, 43 (2002): 63.

[75] The point is to preserve the original data so that they are not potentially modified during PM analysis. Such extraction can also be done by the staff operating PM, as will be described in the text, simply by downloading a local copy of the master database when it needs to use the data. The original master database remains unaffected. (Manager, Medical Informatics, Pacificare, telephone interview with author, May 12, 2004).

[76] Pacific Business Group on Health, Appendix, 5.

[77] "Predictive Modeling, Integrated Disease Management Emerge as Popular Strategies," *Data Strategies and Benchmarks*, 6 (2002).

[78] Welch, 359.

[79] PM, and the other BQMA, currently use identifiable PHI. (See American Association of Health Plans, "Statement on the Confidentiality of Medical Information and the Medical Information Protection Act of 1998"; Government Accounting Office, "Medical Records Privacy: Access Needed for Health Research, But Oversight of Privacy Protections Limited," February 1999, <http://www.gao.gov/archive/1999/he99055.pdf> (3 April 2005); America's Health Insurance Plans,

be examined to add new applications or refine PM techniques.[80] PM is often managed by a staff that performs such tasks. Identifiable data may also be needed for quality control. The claims data may have errors in them. Data entry mistakes may create incorrect values.[81] Software upgrades may lead to bad data formats. Sometimes errors are due to the financial arrangements under which the insurer operates.[82] Under "capitated" arrangements, the insurer pays providers a flat fee per enrollee per month.[83] In this environment, claims may not be submitted for direct provider reimbursement but mostly for administrative or management purposes. Therefore, extensive adjudication, that is, error cleaning and resolution in claims data for reimbursement purposes, might not get done and remaining data may be more prone to error.[84] PM staff cleans some errors.[85] However, it cannot identify all errors. The claim record contains the patient's name, diagnosis, length of stay in a hospital, and other sensitive data, as will be shown later on.[86] [87]

People have concerns about the misuse of identifiable data within organizations. Concerns exist about organizations insufficiently protecting data from outside "hackers" as well as from internal employees. In 2005, almost 50.5 *million* records on individuals and families have been exposed in the US due to lax organizational IT security practices. Personal data in a variety of organizations, including health care, have been subject to theft, hacking, and poor data transmission.[88] Health care organizations have been subject to such faults over the past decade specifically.[89] The insurer should be aware of problems stemming from internal PHI abuse in particular.[90] Across all industries, including health care, it is important to protect IT assets from the "insider threat," a threat from a regular or contract organizational employee who misuses the information he is authorized to use, or, based on his knowledge of the organization's operations, information to which he should have

"Personal Health Plan Information, Health Plans, and Consumers," *AHIP Center for Policy and Research*, August 2001. <http://www.ahipresearch.org/PDFs/24_CRAfinalreportPriv-Conf.pdf> (25 August 2005)).

[80] Implied, Pacificare, Medical Informatics staff, telephone interview with author. August 1, 2003.

[81] See similar concepts in A. J. Dalrymple, L. S. Lahti, L. J. Hutchison, and J. J. O'Doherty, "Record Linkage in a Regional Mental Health Planning Study: Accuracy of Unique Identifiers, Reliability of Sociodemographics, and Estimating Identifier Error," *Journal of Mental Health Administration*, 21 (1994): 187-8.

[82] For example, see Capitation Management Report, "Are You Ready to Take on Claims Adjudication?" September 1999, <http://www.phoenixservice.net/Articles/article6.pdf> (3 September 2005).

[83] W.K. Kellogg Foundation, 11.

[84] HealthcareIndustryPulse, "Payment Errors Cost MCOs Big Money," January 2005, <http://www.bdo.com/about/publications/industry/hcp_jan_05/claims.asp> (4 September 2005).

[85] IS Manager, Tufts Health Plan, telephone interview with author, November 12, 2004.

[86] See Centers for Medicare and Medicaid services, "Health Insurance Claim Form" (HCFA 1500), <http://cms.hhs.gov/providers/edi/cms1500.pdf> (10 October 2003).

[87] Centers for Medicare and Medicaid Services, "Uniform Bill" (UB92), <http://cms.hhs.gov/providers/edi/h1450.pdf> (10 October 2003).

[88] Privacy Rights Clearinghouse, "A Chronology of Data Breaches Reported Since the ChoicePoint Incident," 30 August 2005, <http://www.privacyrights.org/ar/ChronDataBreaches.htm> (2 September 2005).

[89] HIPAAps Privacy and Security.

[90] Deborah Radcliff, "Invisible Loot," *Industry Week*, 2 November 1998, <http://www.industryweek.com/CurrentArticles/ASP/articles.asp?ArticleId=298> (24 September 2005).

no access.[91] [92] [93] A 2000 analysis mentions that the Federal Bureau of Investigations Computer Crime Unit reported that over 80% of network security breaches are "inside jobs" by disgruntled or dishonest employees.[94] HIPAA itself encourages the use of deidentified data whenever it is possible within health care.[95]

Concerns about errors exist, too, which is another privacy concern, as item 8 the Data Quality tenet from the Ethical Force Program list of privacy protections, demonstrates. In this thesis, we focus on errors in identifiers such as the medical record number, used for linking records in the BQMA.[96] [97] Such errors can have clinical impact. In the case of PM, individuals may not be identified or may be improperly risk-stratified and receive improper disease management services, lessening care enhancement effects. One company that has worked on many master patient indices—efforts to accurately merge patient-level data despite potentially lack of consistent identifiers within health organizations—estimates a medical record number "duplication" rate of 10%.[98] [99] Duplication is defined as multiple medical record numbers assigned to the same patient or two or more different patients assigned the same medical record number. One of the company's studies discusses patient clinical decline that can take place as a result of errors in identifiers.[100]

The same information processing exists for the other BQMA. Therefore, those applications may be subject to the same privacy concerns. The other BQMA often have their own staff specializing in their functionality, too.[101] All the BQMA may even be associated with the same claims data set, one copy data store with identifiable PHI driving all the BQMA.[102]

The insurer provides general privacy protection for all data within the organization. For example, it follows HIPAA. According to a 2005 survey, a number of HIPAA tenets are obviously in place, with most health plans distributing a Notice of Privacy Practices and

[91] David Katz, "Elements of a Comprehensive Security Solution," *Health Management Technology*, 21 (2000): 12.
[92] Radcliff.
[93] See National Research Council, 59-60.
[94] Katz, 12.
[95] US Department of Health and Human Services (HHS), "Standards for Privacy of Individually Identifiable Health Information" (part 2), 28 December 2000, 82543, <http://www.hhs.gov/ocr/part2.pdf> (10 October 2003).
[96] See similar concepts in AJ Dalrymple, 187-8.
[97] Lisa I. Iezzoni, *Risk Adjustment for Measuring Healthcare Outcomes, second edition* (Chicago, IL: Health Administration Press, 1997), 224.
[98] Healthcare Informatics Online, "Will Your Patient Data Merge With You?" 1997, <http://www.healthcare-informatics.com/issues/1997/04_97/merge.htm> (31 March 2005).
[99] Lorraine Fernandes, Celia Lenson, Joe Hewitt, Jerry Weber, and Jo Ann Yamamoto, "Medical Record Number Errors," White Paper from Initiate Corporation, April 2001, 3.
[100] Fernandes, 3.
[101] Landacorp staff, telephone interview with author, October 16, 2003.
[102] Landacorp staff, telephone interview with author, October 16, 2003.

obtaining patient authorizations for use and disclosure of PHI.[103] Some errors are cleaned, as discussed earlier. Yet privacy concerns continue. Why doesn't the insurer offer more privacy protection than it currently provides?

1.4 Factors in Technology Adoption by Organizations

We construct a technology adoption model to understand how the insurer might view BQMA privacy protection. Reviewing the literature on organizational technology adoption, we discern four criteria by which an organization may adopt technology. In this discussion, the word "technology" means any technical or procedural mechanism, not only technical, used to improve organizational operations. Organizations will adopt technology due to: 1) the external environment—the technology is required in response to external pressure such as standards dictated by a parent organization or pressure from the community in which the organization operates; 2) economic efficiency—the technology will improve the organization's financial position; 3) organizational context—the technology aligns with the "reputation" the organization wants, its mission, the desires of key employees, and similar organizational factors; and 4) technical efficiency—the technology possesses superior technical characteristics as compared to currently used approaches.[104] [105] [106] [107] [108] [109]

[103] Healthcare Information and Management Systems Society, "U.S. Healthcare Industry: HIPAA Compliance Survey Results: Winter 2005," <http://www.himss.org/Content/files/WinterSurvey2005.pdf> (25 August 2005).

[104] Vivian Carpenter and Ehsan H. Feroz, "Institutional Theory and Accounting Rule Choice: An Analysis of Four US State Governments' Decisions to Adopt Generally Accepted Accounting Principles," *Accounting, Organizations, and Society*, 26 (2001): 571.

[105] Paul Jen-Hwa Hu, Patrick Chau, and Olivia Liu Sheng, "Investigation of Factors Affecting Healthcare Organization's Adoption of Telemedicine Technology" (Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000), 2, 4.

[106] Md. Mahbubur Rahim, G. Shanks and R.B. Johnston, "Understanding Motivations for IOS Adoption" (Proceedings of the Twelfth Australasian Conference on Information Systems).

[107] A. Zutshi and A. Sohal, "Environmental Management System Adoption by Australasian Organizations: Part 1: Reasons, Benefits and Impediments," *Technovation*, 24 (2000): 342.

[108] Rand Corporation, "How MCO Medical Directors See the System," *Managed Care and the Evaluation and Adoption of Emerging Medical Technologies*, 2000, 33, <http://www.rand.org/publications/MR/MR1195/MR1195.chap4.pdf> (31 March 2005).

[109] Rand Corporation, "How Might Technology Adoption be Improved," *Managed Care and the Evaluation and Adoption of Emerging Medical Technologies*, 2000, 49, <http://www.rand.org/publications/MR/MR1195/MR1195.chap6.pdf> (27 March 2005).

1.5 *Insufficient Support for Adoption of Privacy-protecting Practices*

If technology adoption data regarding privacy-preserving practices are analyzed without rigor, little support is found for adding additional privacy protections to the BQMA.

1.5.1 Legal Analysis

We find that the external environment does not encourage BQMA privacy protection beyond what the insurer currently provides. We focus on the external driver of regulation.[110] [111] The insurer, like all US organizations, must abide by appropriate federal and state requirements.[112] [113] [114] Any organization might not be viable otherwise, due to the costs of litigation or potential criminal prosecution. HIPAA is the main regulation affecting health privacy practices in the US. HIPAA does not require extended BQMA privacy protection. HIPAA defines which entities are covered by its provisions and what kinds of data are to be protected:

A) Health plans, health care providers, and health care clearinghouses are the primary "covered entities" that must abide by HIPAA.[115]

B) Protected Health Information (PHI) is individually identifiable health information relating to a physical or mental health condition of an individual, the provision of his care, or the payment for that care.[116]

C) If PHI is de-identified such that it is impossible to identify the data subjects in the data, HIPAA tenets do not apply.

HIPAA requires the following health privacy practices from covered entities:[117] [118]

1) The Consent requirement for Treatment, Payment, and Operations (TPO) was made optional in the latest version of HIPAA.[119] Many individuals would like

[110] Implied, Md. Mahbubur Rahim, "Understanding Motivations for IOS Adoption."

[111] S.R. Elliot, "Adoption and Implementation of IT: An Evaluation of the Applicability of Western Strategic Models to Chinese Firms," in *Diffusion and Adoption of Information Technology*, ed. Karlheinz Kautz (London: Chapman & Hall, 1996), 18.

[112] For example, see General Accounting Office, "Health Insurance Regulation: Varying State Requirements Affect Cost of Insurance," 1999, <http://www.gao.gov/archive/1996/he96161.pdf> (22 March 2005).

[113] Roland Strum and J. Unutzer, "State Legislation and the Use of Complementary and Alternative Medicine," *Inquiry – Blue Cross and Blue Shield Association*, Winter 2000/2001, 425-6.

[114] Managed Care Magazine, "State Mandates Promote Contraceptive Coverage," 2004, <http://www.managedcaremag.com/archives/0408/0408.formularyfiles.html> (18 March 2005).

[115] A "health care clearinghouse" is a public or private entity that transforms nonstandard data or health care-related transactions received from another entity into "standard" transactions or data elements. (See Centers for Disease Control and Prevention, "HIPAA Privacy Rule and Public Health").

[116] As was defined in the beginning of the thesis.

[117] Taken from Centers for Disease Control and Prevention, "HIPAA Privacy Rule and Public Health."

[118] US Department of Health and Human Services, "Protecting the Privacy of Patients' Health Information: Summary of the Final Regulation," 2000, <http://www.hhs.gov/news/press/2000pres/00fsprivacy.html> (18 October 2003).

[119] HHS (update), 53208-10.

organizations in the health industry to ask for consent before they disclose PHI. However, many health care providers indicate that consent should not be required for payment, treatment, or routine organizational operations as it might impair care delivery.[120] Providers offering direct care to patients are obliged to make a good faith attempt to receive a patient's acknowledgement of receipt of Notice as will be described in point 3 below. The patient may also request restrictions to PHI use as will be discussed in point 6 below.[121] [122] Ultimately, however, PHI may be used for TPO without difficulty.

2) Authorizations are required, with some exceptions, for several types of data handling functions not explicitly permitted by HIPAA.[123] For example, using PHI for marketing purposes requires an individual's authorization.[124]
3) "Covered entities" must provide:
    b. Notice. Individuals must be given notice describing their privacy rights and how their PHI will be used or disclosed.
    c. Access. Individuals have a right to access their health information, including certain rights to amend their PHI.
    d. Security. The covered entity must have in place appropriate administrative, technical, and physical safeguards to protect the privacy of health information.
4) Some of the covered entities' staff must be designated to implement the organization's privacy practices and receive complaints about them.
5) Individuals may request an "accounting" of some PHI disclosures by the covered entity for certain transactions, except for TPO and some other transactions.[125] [126]
6) Individuals may request restrictions on use of PHI. These requests can be directed toward TPO functions.[127] However, the covered entity is not obligated to agree to such requests.
7) Covered entities may disclose PHI without authorization when required by law, for public health, or for other special reasons.
8) Use or disclosure of information by the covered entity is limited to the minimum necessary for the work associated with the use or disclosure.
9) Civil and criminal penalties are prescribed for various violations, the most egregious of which carries a penalty of $250,000 plus 10 years in jail if the intent was to sell PHI for personal gain.
10) Covered entities that conduct business with "business associates," such as third-party claims processors, external consultants and auditors, and lawyers, and

[120] HHS (update), 53209.

[121] University of Miami, "Privacy/Data Protection Project," 15 August 2002, <http://privacy.med.miami.edu/glossary/xd_consent.htm> (7 April 2005).

[122] American Academy of Ophthalmic Executives, "Final HIPAA Privacy Rule," <http://www.aao.org/aaoesite/promo/compliance/hipaa_final.cfm> (9 April 2005).

[123] HHS (update), 53220.

[124] HHS (update), 53220.

[125] HHS (part 1), 82559-82560.

[126] HHS (update), 53243-53245.

[127] See HHS (update), 53211.

transfer PHI to them must sign contracts with them with privacy protection requirements.[128] [129]

11) Stronger state privacy laws continue to apply as HIPAA just provides a federal baseline standard.

1.5.1.1 Basic Quality Management Applications Under HIPAA

The BQMA copy data store(s) uses identifiable data, item *(C)* above. We again focus on PM for our privacy analysis and extend the results to the other BQMA. PM is subject to HIPAA compliance points 1 through 11 due to identifiable PHI use. Tenets 1 through 5 do not prevent the copy data store from being identifiable, thus potentially leading to the information abuses discussed before. Legally, the PM part of disease management is considered Treatment, Payment, and Operations (TPO). Analyzing claims to find individuals who could benefit from goods or services designed to improve health care or reduce cost, the PM part of disease management, is considered health care operations.[130] Per point 1, a covered entity may optionally ask for consent. We assume the insurer will not ask for it regarding PM because it may be financially disadvantageous. Using identifiable PHI for PM may be profitable to the insurer because disease management might reduce the costs of policy-holder care. If the insurer asks for consent and policy-holders in some way prevent PM from using identifiable data, the insurer's profitability might be impacted, hence minimizing its interest in consent. Point 2 indicates that for several types of data handling functions not permitted by HIPAA, the covered entity must seek authorization. However, since the usage of PM falls under TPO, an allowed data practice, the need for an authorization is bypassed. Point 3 requires a covered entity to give Notice. However, the Notice is in no way a consent mechanism. Policy-holders have no opportunity to agree or disagree with the process. Point 3 also requires a covered entity to provide Security. In the PM context, a person must not be able to see PHI inadvertently.[131] This requirement alone does not prevent identifiable PHI use. Identifiable data are needed to operate and update the PM application, as explained before. The insurer must provide general security surrounding PHI. However, once employees who run the PM platform access the data legitimately they are allowed full access to those data, in this case identifiable PHI. Point 5, an "accounting" of disclosures made, does not apply because PM is part of TPO and TPO functions are exempted by this tenet.

[128] See University of Texas Health Science Center at San Antonio, "Evaluation for Business Associates," *HIPAA Compliance Program*, 10 October 2005, <http://www.uthscsa.edu/hipaa/assoc-who.html> (23 December 2005).

[129] US Department of Health and Human Services, Office for Civil Rights, "Business Associates," 3 April 2003, <http://www.hhs.gov/ocr/hipaa/guidelines/businessassociates.rtf> (23 December 2005).

[130] For example, see Atlantic Information Services, "HIPAA Compliance Strategies," 2003, <http://www.aishealth.com/Compliance/Hipaa/MCWDMTraining.html> (13 October 2003).

[131] US Department of Health and Human Services (HHS), "Standards for Privacy of Individually Identifiable Health Information" (part 3), 28 December 2000, 82561-2, <http://www.hhs.gov/ocr/part3.pdf> (13 October 2003).

Tenets 6 through 11 also do not prevent identifiable PHI use, again allowing for potential information mishandling. Point 6 allows for individuals to restrict how PHI is used or disclosed. Even if the individual specifies desired restrictions the covered entity is not obligated to agree. Again, using identifiable PHI for PM may be profitable, therefore, agreeing to any PHI restrictions may not be in the insurer's interest. HIPAA's point 8 the minimal use requirement does not apply either. The covered entity must limit data access in a manner consistent with the employee's need.[132] The classes of people who need PHI, the types of PHI, and the conditions appropriate to the access must be understood. "Reasonable" determinations should be made to restrict PHI access consistent with a user's job. PM staff can run and modify the PM application. As implied from a conversation with staff of the US Department of Health and Human Services Office for Civil Rights which enforces HIPAA, granting identifiable PHI access to PM staff for such functions should be permitted.[133] HIPAA allows for even stricter state law to take precedence. However, in general, state statues do not support a strong notion of consent when it comes to using health information internal to an entity.[134] Per state law, it appears that internal information use is even less constricted.

Similar analysis can be made for the other BQMA. They can also be shown to permit the usage of identifiable data, as they are also part of TPO.[135] [136]

1.5.2 Financial Perspective

Providing extended BQMA privacy protection also does not appear profitable, another important component of the technology adoption model. We use a decision analytic framework to demonstrate financial implications. Operationally, decision analysis uses the values and perceived uncertainties of decision makers to choose an action providing maximal expected value to the decision makers.[137] [138] We model the making of a financial decision per the descriptions in Thompson, Barr, and Hunink:[139] [140] [141]

[132] HHS (part 2), 82544.

[133] Implied, HHS, Office for Civil Rights staff, telephone interview with author, September 25, 2003.

[134] HHS (part 1), 82473.

[135] See for example Alcohol, Drug, and Mental Health Board for Franklin County, "Info for Consumers," <http://www.adamhfranklin.org/consumers/hipaaPolicy05.php> (19 April 2005).

[136] Jack A. Rovner, "Don't Let Fear of HIPAA Keep You from Crucial Data," *Managed Care Magazine*, March 2003, <http://www.managedcaremag.com/archives/0303/0303.legal.html> (5 April 2005).

[137] Mark S. Thompson, *Decision Analysis for Program Evaluation* (Cambridge, MA: Ballinger Publishing Company, 1982), 8.

[138] Judith Barr and Gerald Schumacher, "Using Decision Analysis to Conduct Pharmacoeconomic Studies," in *Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition,* ed. B. Spilker, 1198 (Philadelphia: Lippincott-raven Publishers, 1996).

[139] Mark S. Thompson, 11-12.

[140] Barr, 1197-1214.

[141] Myriam Hunink and Paul Galsziou, *Decision Making in Health and Medicine: Integrating Evidence and Values* (Cambridge, UK: Cambridge University Press, 2001), 251-266.

1) Determine from which point of view the decision should be made, such as a patient's, a company's, or society's. Costs and benefits will be modeled from this perspective.
2) Identify the overall decision, including the relevant timeframe.[142]
3) For each decision path, a choice to make within the decision, structure the actions and associated consequences over time.[143]
4) Assess the probability and economic outcome of each consequence in each decision path.[144] [145]
5) Combine the probabilities and associated magnitudes of all events to arrive at a final expected value for each path.
6) Select the path with the greatest expected value.
7) Conduct a sensitivity analysis to determine decision robustness based on underlying financial parameters.[146]

We apply the framework to BQMA privacy protection. The time frame considered for step 2, regarding whether additional BQMA privacy protections should be adopted, will be 12 months. This is a relatively short time frame. We may assume the insurer will be focused on decisions with short-term impacts because it might better control such decisions.

Steps 4 and 5 require quantifying and combining the probabilities and magnitudes along each decision path. We explore if available data permit quantifying the gains and losses of adopting versus not adopting extra BQMA privacy protections.

1.5.2.1 Difficulty in Measuring Privacy Benefits

Unfortunately, it's difficult to quantify the benefits of providing privacy protection. Some benefits are hard to uncover. Avoiding litigation would be a key financial benefit; the insurer faces fewer lawsuits. Yet enforcement under HIPAA has been lax. Health organizations have less to fear financially because they can address breaches and face limited financial repercussions. HIPAA does not provide for a private cause of action.[147] Consumers must complain to the US Department of Health and Human Services (HHS), which will investigate their complaints. Because consumers may not recognize privacy violations and are not part of health organizations to understand how health information might be misused, the number of complaints may not be large.[148] Furthermore, the intent

[142] Barr, 1203.
[143] Barr, 1205.
[144] Barr, 1207-8.
[145] See for example Hunink, 40.
[146] Hunink, 19.
[147] Wiley Rein & Fielding LLP, "A New Era for HIPAA Enforcement," May 2004, <http://www.wrf.com/publication_newsletters.cfm?sp=newsletter&year=2004&ID=10&publication_id=9825&keyword> (29 August 2005).
[148] iHealthbeat, "Enforcement of HIPAA Privacy: Making it Real," 19 November 2003, <http://ihealthbeat.org/index.cfm?Action=dspItem&itemID=100262> (29 August 2005).

of HHS' enforcement approach is to seek voluntary compliance from covered entities. Punishment may only come if voluntary reconciliation is ineffective.

The benefits of incorporating privacy-protecting policies and technologies may be categorized. The losses stemming from inadequate privacy protection may be delineated into measurable loss categories despite the "intangible" nature of privacy.[149] [150] [151] [152] How customers purchase fewer goods or services from an organization; how customers recommend an organization to others less often; or how an organization cannot acquire as many new customers over time as before could be the measurable loss categories.[153] [154] [155] A measure can be obtained for each category. Implementing stronger privacy protections would reverse such losses, quantifying the *benefits* of stronger privacy protections.

Unfortunately, the data needed for such categories are often unavailable, and available data are not robust. The biggest problem is simply data availability. Several commentators have pointed out that there are disincentives for organizations to publish information about their security breaches, which is needed for quantifying the losses described above.[156] [157] The organization may face legal liability as customers become aware of how their data have been abused, or loss of reputation among business partners who perceive lax security practices. The data that are available are not gathered based on sound research principles. Worries regarding the publication of security data and other reasons skew available statistics. The CSI/FBI survey is a popular annual survey on IT abuse. Among other features it contains financial loss categories into which respondents may place their financial losses for IT security for that year. Sabotage, System Penetration, and Net Abuse are some of such categories.[158] The 2002 CSI/FBI survey confirms that over the years, the response rate to the CSI/FBI survey has been low. Out of the approximately 3500-4000 annual questionnaires mailed out, the response rate has been about 14% from 1999-2002.[159] Using the loss categories from the survey may lead

[149] G. Stevenson Smith, "Recognizing and Preparing Loss Estimates from Cyber-Attacks," *Information Systems Security*, 12 (2004); 47.

[150] Fernandes, 5-6.

[151] Kevin J. Soo Hoo, "How Much is Enough? A Risk-management Approach to Computer Security," June 2000, 40-41, <http://iis-db.stanford.edu/pubs/11900/soohoo.pdf> (12 September 2005).

[152] Health Privacy Project, "Best Principles for Health Privacy," July 1999, <http://www.healthprivacy.org/usr_doc/33807.pdf> (Apr 1, 2005).

[153] Chang Liu, Jack T. Marchewka, June Lu, and Chun-Sheng Yu, "Beyond Concern: A Privacy-Trust-Behavioral Intention Model of Electronic Commerce," *Information & Management*, 42 (2004):135-6.

[154] Steve Kenny and John Borking, "The Value of Privacy Engineering," *Journal of Information, Law, and Technology*, 1 (2002), <http://www2.warwick.ac.uk/fac/soc/law/elj/jilt/2002_1/kenny/> (1 April 2005).

[155] Bernd W. Wirtz and Nikolai Lihotzky, "Customer Retention Management in the B2C Electronic Business," *Long Range Planning*, 36 (2003): 519.

[156] Computer Security Institute, "2004 CSI/FBI Computer Crime and Security Survey," 2004, <http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2004.pdf> (1 April 2005).

[157] Kevin J. Soo Hoo, 31.

[158] See Computer Security Institute, "2001 CSI/FBI Computer Crime and Security Survey," Spring 2001, <http://www.reddshell.com/docs/csi_fbi_2001.pdf> (1 April 2005).

[159] Computer Security Institute, "2002 CSI/FBI Computer Crime and Security Survey," 2002, <http://www.reddshell.com/docs/csi_fbi_2002.pdf> (1 April 2005).

to bias because actual losses could be higher or lower than published results due to significant *non-response.*

Another problem is the difficulty in measuring losses *solely* due to privacy. For example, the US Post Office rents out 18 million post office boxes (PO boxes) for $500 million per year.[160] The US Post Office explicitly lists privacy as one motivator for such rental. Even if one can quantify a person's "willingness to pay" (WTP) for such *privacy*—that is, quantify people's valuation of privacy based on how much they're willing to pay for mailbox rental—professional appearance may be another reason why people rent PO boxes.[161] The just computed WTP value would have to be further divided into a "privacy" WTP and a "professionalism" (or something similar) WTP so that one can extract the privacy-specific WTP.

Of course to avoid the difficulties with published studies, the organization can carry out its own research. It can quantify its own measures of privacy protection. However, this endeavor would be complex, too, requiring robust social research. HIPAA and other federal and state laws have been passed. Consumer surveys continue to show people's desire for medical privacy.[162] [163] Yet there appear only few current valuations of privacy protection, medical or otherwise, regarding costs and especially benefits. In 2003, the federal government's Office of Management and Budget (OMB) asked experts around the country how to measure the "costs" of potential civil liberties and privacy intrusions in the Bush Administration's push for tighter domestic security for better counter-terrorism.[164] How can one value lost time, lost privacy, and similar concepts so that the price of increased security can be better ascertained and compared with "benefits"? The OBM acknowledged that the end results may not necessarily even be quantifiable in dollars.

The challenge in valuation, to which OMB's request points, is that the valuation effort might not be simple. From the definition of privacy earlier, privacy protection is ultimately defined by the consumer. Her perspective should be sought regarding whether privacy protection is properly provided. The Ethical Force Program (EFP) guidelines earlier underpin such a perspective. Policies promulgated to protect privacy within organizations are inconsistent, limiting basic understanding of privacy safeguards provision. The EFP guidelines mention that a variety of Fair Information Practices-type of definitions exist in the US and the world.[165] The Joint Commission on Accreditation of Healthcare Organizations and the National Committee for Quality Assurance which create accreditation standards for health organizations; the American Bar Association; and the HIPAA statute itself all suggest somewhat different ways of instantiating privacy within organizations. Consumers may not have a consistent notion of what is "privacy."

[160] Adam Shostak, "'People Won't Pay for Privacy,' Reconsidered," 14 March 2003, <http://www.cpppe.umd.edu/rhsmith3/papers/Final_session3_shostack_privacy.pdf> (1 April 2005).
[161] Shostak.
[162] Health Privacy Project, "Health Privacy Polling Data."
[163] Harris Interactive, "Privacy On and Off the Internet: What Consumers Want."
[164] See James Love, "NYT on OMB's Costs Benefit of Losses of Liberty and Privacy," 2 April 2003, <http://lists.essential.org/pipermail/random-bits/2003-April/001054.html> (8 September 2005).
[165] American Medical Association, "The Ethical Force Program," 10-12.

Further, as mentioned before, many consumers might not know the nature of PHI use within health organizations.[166] Health industry terms, specific organizations, and the health industry itself might have to be explicated to consumers to solicit specific privacy valuations.

Such work may take time and could lead to error. Smith et al. spent four years, from late 1989 to late 1993, constructing a privacy survey instrument measuring individuals' attitudes towards organizational privacy practices.[167] They interviewed hundreds of people across the US, computing standard survey internal and external validity metrics.[168] Their focus is not completely appropriate for this thesis as they focused on somewhat broader organizational issues. Nevertheless, what is relevant is that Stewart et al. administered this survey in 2002. Stewart et al. showed that consumers may have additional privacy concerns beyond those that Smith and his colleagues originally thought.[169] Smith originally posited that people are concerned about too much data collection, unauthorized secondary use of data, improper access to data, and that data errors are not sufficiently cleaned.[170] Stewart found that this is true. However, consumers may have additional information control concerns. They may want to exercise more direct control over their data, such as getting access to them or being asked permission to collect their personal information from the collecting organization.[171] Can the insurer handle the associated methodological issues?

1.5.2.2 Cost of Solutions for Privacy Protection

The costs of adopting privacy-enhancing technologies and policies are easier to estimate than their benefits. To address the BQMA privacy concerns, data must be deidentified, errors must be handled in linkage identifiers, and ultimately data must be reidentified in some cases, such as for disease management, so that staff can follow up with policy-holders as needed. The main approach explored in this thesis is new deidentification techniques that allow the BQMA to function, potentially, without loss of performance. We will discuss the techniques in the technical part of this thesis. Outside of such technical solutions, however, as we will see, there are few current solutions to provide needed BQMA privacy protections. One solution to privacy concerns raised by BQMA is the simple but extreme alternative of terminating the applications. Identifiable data would no longer be used. However, the BQMA may currently save the insurer money; therefore this solution is clearly not viable. An extended review of privacy-protecting solutions

[166] C. Shawn Tracy, Guilherme Coelho Dantas, and Ross EG Upshur, "Feasibility of a Patient Decision Aid Regarding Disclosure of Personal Health Information: Qualitative Evaluation of the Health Care Information Directive," *BMC Medical Informatics and Decision Making*, 3-4 (2004), <http://www.pubmedcentral.gov/picrender.fcgi?artid=518970&blobtype=pdf> (29 August 2005).

[167] H. Jeff Smith, 167-196.

[168] H. Jeff Smith, 175-9.

[169] Kathy A. Stewart and Albert H. Segars, "An Empirical Examination of the Concern for Information Privacy Instrument," *Information Systems Research*, 13 (2002): 46.

[170] H. Jeff Smith, 172, 181.

[171] Stewart, 40, 44-45.

shows few other possibilities. Another idea would be to "wait" for the problem to dissolve. PHI concerns may be time sensitive, such as being dependent on family context or care setting.[172] One survey shows that those who spend time online and those who have more rather than less online experience have fewer privacy concerns than non- or new Internet users.[173] As individuals become more comfortable with computer technology and perceive how heath information might be used within health care, they might trust current health data management practices. Over time, privacy concerns regarding an insurer might lessen.

In practice, such a solution would also be unworkable. Concerns about privacy have not declined.[174] Some surveys show some associated levels of health privacy concern over time. In 1993, 85% of the respondents to a Louis Harris and Associates survey said that protecting the confidentiality of medical records was "absolutely essential" or "very important."[175] In a 2000 MedicAlert Foundation survey, 77% of the respondents stated that the privacy of their health information is very important.[176] In early 2003, I informally surveyed eight genetic counselors who dealt with the rare genetic disorder Huntington's disease. The counselors said between 20-80% of their patients getting tested for the Huntington mutation paid for the test out of pocket due to privacy fears.[177]

Other surveys imply an increase in privacy concern. A 2002 Harris Interactive survey indicated that people may be divided into three categories regarding privacy protection.[178] The privacy "fundamentalists" are those who feel privacy protection is a core right and many organizations should generally not get the personal information they seek. The privacy "pragmatists" are those who weigh the potential benefits provided by organizations against costs of supplying personal information. The privacy "unconcerned" are those who don't care much about privacy safeguards and more willingly provide personal information despite warnings of potential privacy abuse. From the second half of the 1990s to 2001, the percent of people self-identifying as "fundamentalists" went up from 25% to 34% while the percent self-identifying as "unconcerned" went down from 20% to 8%. Indeed, privacy concerns might be increasing.

The other privacy-protecting solutions in the literature are *prevention* technology-based solutions, which are mechanisms to prevent individuals from accessing data.[179] [180] [181]

[172] Health Privacy Project, "Best Principles for Health Privacy."

[173] The Pew Internet & American Life Project, "Trust and Privacy Online: Why Americans Want to Rewrite the Rules," 20 August 2000, 3, <http://www.pewinternet.org/pdfs/PIP_Trust_Privacy_Report.pdf> (30 August 30, 2005).

[174] Harris Interactive, "Privacy On and Off the Internet: What Consumers Want."

[175] Center for Democracy and Technology, "Statement of Janlori Goldman, House Committee on Government Reform and Oversight."

[176] Health Privacy Project, "Health Privacy Polling Data."

[177] Interviews with 8 Huntington's disease genetic counselors nationwide. Conducted February 24, 2003 - March 8, 2003.

[178] Harris Interactive, "Privacy On and Off the Internet: What Consumers Want," 19-22.

[179] See for example National Research Council, 61.

These solutions include encryption, data access controls, or query or output restrictions. We will discuss these solutions, including the identifier error problem, in the technical part of the thesis.

However, what is clear is that any solution will have a cost. Procedures or technical upgrades will have to be incorporated into employees' workflows. These changes will have a financial impact. Management, operations, and maintenance costs have been the assumed costs of integrating security measures into the IT environment.[182]

Following step 6 in our financial decision model, choosing the best alternative, not providing additional BQMA privacy protection may be financially most advantageous for the insurer. Methodological difficulties and unavailable data hamper the financial valuation of practices that protect privacy. Whether from a 12-month perspective, or even from a longer-term perspective, providing extra BQMA privacy protection does not create a clear return on investment for the insurer because benefits are vague whereas costs are less so. We will show later how a stronger financial case can be made for adopting stronger BQMA privacy protections when data are analyzed more rigorously. However, an unrigorous data analysis does not demonstrate clear financial benefits.

1.5.3 Organizational Perspective

Regarding the organizational context of the technology adoption model, we will focus on the insurer's motivation to improve quality of care. The insurer's environment may be focused on quality of care, encouraging the insurer to focus on the goal. Within the US, health care organizations are encouraged to offer quality care. Consumers; the government, including a President's Commission; accreditation organizations; and national organizations that monitor US care quality all promote US health care quality.[183] [184] [185] [186] [187] Health insurance organizations may also have such a goal.[188] For example,

[180] Implied, Federal Committee on Statistical Methodology Confidentiality and Data Access Committee, "Restricted Access Procedures," 4 April 2002, 8, <http://www.fcsm.gov/committees/cdac/cdacra9.pdf> (30 August 2005).

[181] Carol A. Siegel, Ty Sagalow, and Paul Serritella, "Cyber-risk Management: Technical and Insurance Controls for Enterprise-level Security," *Information Systems Security*, 11 (2002): 40, 45, 48.

[182] Thomas C. Rindfleisch, "Privacy, Information Technology, and Health Care," *Communications of the ACM*, 40 (1997): 99.

[183] See Leonard Friedman and D.B. White, "What Is Quality, Who Wants It, and Why?" *Managed Care Quarterly*, 7 (1999): 43.

[184] Joint Commission on Accreditation of Healthcare Organizations, "What is the Joint Commission on Accreditation of Healthcare Organizations," 2005, <http://www.jcaho.org/general+public/who+jc/index.htm> (30 August 2005).

[185] Agency for Healthcare Research and Quality (AHRQ), "What is AHRQ?" February 2002, <http://www.ahrq.gov/about/whatis.htm> (30 August 2005).

[186] Karen Sandrick, "Tops in Quality," *Trustee*, 56 (2003): 14.

[187] President's Advisory Commission on Consumer Protection and Quality in the Health Care Industry, <http://www.hcqualitycommission.gov/> (30 August 2005).

[188] Rand Corporation, "How MCO Medical Directors See the System."

one study showed that if an HMO was profitable in a prior period, its quality of care metrics were improved in a subsequent period. As the study concluded, HMO profitability may allow the HMO to invest its resources in improving services which in turn might enhance care.[189]

From a quality of care perspective, problems similar to those for identifying financial benefits exist in identifying how extra BQMA privacy protection may improve care. Few studies assess the impact of using privacy-protecting policies and technologies on care provision, lessening demand for such protections. The company that created the master patient indices discussed earlier indicates that many organizations may not be aware of the care delivery impact of identifier errors.[190] [191] The degree of suboptimal PM linkage or the reduction in medical effectiveness due to poor linkage within software applications may be unclear to organizations.

1.5.4 Technical Perspective

Technical efficiency in the context of our technology adoption model will mean utilizing a more secure and efficient technology in providing privacy protection. As we mentioned before, there are apparently few current technical solutions that meet the requirements of BQMA privacy protections. We will create new technology to provide such protections later on. However, without new approaches few other solutions appear to exist, lessening demand for needed protections.

1.6 Chapter Conclusion

A less rigorous analysis of data does not demonstrate the benefits to the insurer of adding additional BQMA privacy protections. From a regulatory, economic, organizational, and technical perspective available data do not support adopting such protections. Federal law does not require them; such protections do not appear to provide financial or quality of care benefits; and technically they appear difficult to create. A more convincing argument can be made for adopting such protections when analyzing existing and new data in more depth. We conduct such an analysis in chapters two and three.

[189] Patricia H. Born and Carol Simon, "Patients and Profits: The Relationship between HMO Financial Performance and Quality of Care," *Health Affairs*, 20 (2001): 167-74.
[190] Healthcare Informatics Online. "Will Your Patient Data Merge With You?"
[191] Fernandes, "Medical Record Number Errors," 3, 7, 9.

2 Contextual, Financial, and Organizational Support for Stronger Privacy Protections in Routine Applications

Upon closer inspection, available data support adding extra BQMA privacy protections. In this chapter, we will more closely investigate the legislative, financial, and quality of care benefits. The technical mechanisms to provide such protections will be given in chapter three. The outline of this chapter is as follows. We will first highlight the new laws which are being considered because identifiable data have been significantly recently misused in the US. A new cost model will be subsequently presented demonstrating the detailed financial benefits and costs of adding stronger BQMA privacy protections. Capturing a somewhat realistic competitive health insurance market, the cost model first describes how some policy-holders may switch to a competitive insurance organization which offers extra BQMA privacy protections. Next, the financial benefit from reducing the ability of BQMA staff to misuse the copy data store(s) because it is in identifiable form will be computed. In the third part of the model, we will explain how the insurer should handle the increase in claims liability it will face should it offer stronger BQMA privacy protections. Some policy-holders will stop paying for medical services out-of-pocket because the insurer provides better privacy safeguards, increasing the insurer's claims expenses. We will describe the steps the insurer should take to obtain the funds to pay for such expenses. Fourth, a disease management model will be constructed to show how improving BQMA privacy protections will allow the insurer to enroll candidates into disease management programs in a timely manner. Expenses to the insurer should be reduced because it has a greater opportunity to prevent complications arising regarding its policy-holders' medical conditions. In the fifth part of the cost model, the same disease management model will be invoked to show how reducing errors in linkage identifiers will also allow for timely enrollment of policy-holders into a disease management program and thus again likely reduce the insurer's expenses. Sixth, the costs of data deidentification will be computed. This is the mechanism we will use to provide the stronger BQMA privacy protections, as will be shown in chapter three. Seventh, the last part of the cost model, will be a sensitivity analysis to demonstrate the contextual parameters impacting the insurer's financial benefits from adding better privacy protections. Finally, to demonstrate improvements in quality of care to policy-holders, another component of our technology adoption model, we will again rely on the disease management model. By enrolling more policy-holders into a disease management program the insurer enhances their care. We will quantify the improvement in care that policy-holders experience as a result of the insurer's efforts to improve their privacy.

2.1 Contextual Support for Privacy Protection in Routine Applications

From the insurer's environmental perspective, by voluntarily embracing more BQMA privacy protections the insurer might avoid new administrative burdens stemming from external regulations that might be passed. The privacy aspects of HIPAA as well as other federal and state laws protecting general privacy have been passed due to rising recent and past concerns about the ease of data collection, transmission, and misuse within health care and other organizations in the US.[192] Democratic- and Republican-sponsored bills have been passing through Congress in 2005 to better protect the privacy of identifiable data within different organizations in the wake of the most recent data breaches at ChoicePoint, Bank of America, and LexisNexis.[193] [194] These laws will affect insurance organization practices because insurers utilize such identifiable data. Additional laws, specific to the health industry, may be passed if internal data within health organizations continue to be subject to misapplication. If the insurer adopts stronger BQMA privacy protections it might lessen the opportunity for the passage of such legislation.

The following is a sample of 2005 bills currently passing through Congress, offering a description of bills that might become laws due to poor privacy protections. These bills would affect insurer practices. Bill S.768, the "Comprehensive Identity Theft Prevention Act," prescribes the following actions, out of an extended list, for a "covered person," i.e. any commercial entity, to follow regarding IT security breaches.[195] If the covered person is subject to a breach wherein there is reason to believe the sensitive information taken can be used to reidentify data subjects, the covered person must notify all people who are believed to have been subjects of the breach as well as the Federal Trade Commission (FTC). Further, consumers, upon receiving such a notice, may request that the covered person expunge their sensitive information from the entity's internal records. There is also a prohibition from soliciting Social Security Numbers (SSN) by any individual unless this is necessary for normal business and no other identifying number can be used. S.1408, the "Identity Theft Prevention Act," also requires a covered entity, i.e., any for-profit or nonprofit organization, to notify consumers, the FTC, and all consumer reporting agencies, such as the national credit agencies Experian or Trans Union, if there is a reasonable basis to conclude some identity theft has happened involving the covered entity's data.[196] [197] There is also a prohibition from soliciting SSNs unless there is a

[192] HHS (part 1), 82469

[193] Declan McCullagh, "Congress Edges toward New Privacy Rules," *News.com*, 10 March 2005, <http://news.com.com/Congress+edges+toward+new+privacy+rules/2100-1028_3-5609324.html?tag=nl> (30 August 2005).

[194] Center for Democracy and Technology, "Consumer Privacy Legislation (109th)," <http://www.cdt.org/legislation/0/3/> (30 August 2005).

[195] The Library of Congress, "S. 768," *Thomas*, <http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:s768is.txt.pdf> (26 December 2005).

[196] The Library of Congress, "S. 1408," *Thomas*, <http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:s1408is.txt.pdf> (26 December 2005).

specific use for them for which no other identifier exists. S.1332, the "Personal Data Privacy and Security Act of 2005," also provides for consumer notification when data have been stolen.[198] Credit reporting agencies must again be notified. The business entity that suffered the breach must pay for a monthly credit report and credit monitoring services for each consumer who was notified about the breach for a period of 1 year after the notice was sent to the consumer. A business entity cannot require an individual to provide his SSN as an account number to obtain goods or services after the passage of this Act. Goods or services must be provided if a person does not or cannot supply the SSN. It's important to note that these bills' SSN-usage limitations might make uses of the SSN more challenging for insurers. Per the Government Accountability Office, as we will later see, some insurers use SSNs as the primary policy-holder identifiers. Modifications to insurers' information systems might be necessary to minimize SSN use. All the bills above prescribe penalties if the tenets described above are not appropriately followed.

2.2 Financial Support

2.2.1 "Competitive" Profitability

From an economic point of view, this thesis demonstrates a new analysis to quantify some financial advantages of installing additional BQMA privacy protections. We quantify some intangible benefits of utilizing privacy safeguards, demonstrating positive returns to a health insurer. Note, in the presentation of a cost model below, we present specific averages of values based on particular assumptions. We relax these assumptions in the sensitivity analysis which follows the construction of the basic model. Also, we present our computed results with one extra significant digit. To extract the "final value" of any computed result one only needs to round off the very last non-zero digit of a completed computation.

Consider the following somewhat hypothetical cost model, which is based on the national health insurance market in 2001. According to Kaiser Foundation research, the total number of non-elderly people in the US in 2001 was 247.5 million.[199] Of these, 64.7%, that is, about 160.1 million individuals, received employment-based insurance. In 2001, in this population 60%, or about 96.0 million people, had a choice of at least two health

[197] Federal Trade Commission, "Nation's Big Three Consumer Reporting Agencies Agree To Pay $2.5 Million To Settle FTC Charges of Violating Fair Credit Reporting Act," <http://www.ftc.gov/opa/2000/01/busysignal.htm> (26 December 2005).

[198] The Library of Congress, "S. 1332," *Thomas*, <http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:s1332pcs.txt.pdf> (26 December 2005).

[199] The Henry J. Kaiser Family Foundation. "Health Insurance Coverage in America, 2001 Data Update," January 2003, <http://www.kff.org/uninsured/loader.cfm?url=/commonspot/security/getfile.cfm&PageID=14309> (1 April 2005).

plan options through their workplace.[200] We focus on these individuals, approximately a third of the US population in 2001. Assume such individuals are part of the health insurance "group market" as employers typically purchase insurance for their employees as a group rather than sponsor the employees' purchase of "individual" insurance.[201]

We examine two competing insurers in this marketplace. We make several simplifying assumptions regarding available marketplace data to synthesize our model. According to 2001 research, all states had a number of insurers in the group market serving people in those states.[202] The market penetration of the top three group health insurers varied from 96% to 30%, and on average was 66%. The penetration of the largest group insurer varied from 91% to 11%, and on average was 39%.[203] Comparing the two different market shares suggests that our 96-million person cohort must have been served by at least two different insurance organizations. Building our hypothetical cost model on these data, our top company would have a 39% market share. We can assume the second company would have a 14% market share, approximately half of the 27% difference between the 66% and 39% penetrations. We assume that the second and third largest insurers had roughly equal distributions of the remaining market share, for simplicity. In our hypothetical model, the total market share of the first and second insurance company would be 53%. To simplify the computations, we will assume equal populations across all US states. Thus, in our model, on average the total number of covered employees per state was 96 / 50, or roughly 1.92 million. Often employees chose their health benefits every 12 months.[204] We therefore also assume that the total number of employees annually choosing only one of the two companies mentioned above was 1,920,000 * 0.53 or approximately 1,010,000 individuals. The 39% penetration means the larger insurer enrolled roughly 748,000 policy-holders, while the smaller insurer, at 14% penetration, enrolled roughly 268,000 policy-holders in 2001.

First, we estimate how many people join the insurer because it provides extra BQMA privacy protection. This helps us quantify in our hypothetical model how much of a gain market share an organization might achieve through the intangible "improvement in reputation" relating to privacy protection, one benefit of incorporating extra privacy protection.[205] [206] [207] To obtain a quantitative estimate for our model, we will use data

[200] The Henry J. Kaiser Family Foundation. "Employer Health Benefits, 2001 Annual Survey," 2001, 63, <http://www.kff.org/insurance/loader.cfm?url=/commonspot/security/getfile.cfm&PageID=13836> (1 April 2005). Note, we assume that although the 60% from this survey corresponds to covered *workers* having at least two health plan choices, the 60% is applied to the original 160.1 million which includes children as well. The 160.1 million is the non-elderly population covered by the workplace. From this same Kaiser Family Foundation survey, 56% workers chose to extend coverage to one or more dependents (spouse or children). (The Henry J. Kaiser Family Foundation. "Employer Health Benefits, 2001 Annual Survey," 55). Therefore, we will extend our analysis regarding which individuals have a choice of at least two health plans to children, too.

[201] AcademyHealth, "Mapping State Health Insurance Markets, 2001: Structure and Change," September 2003, <http://www.statecoverage.net/pdf/mapping2001.pdf> (1 April 2005).

[202] AcademyHealth, "Mapping State Health Insurance Markets, 2001: Structure and Change."

[203] AcademyHealth, "Mapping State Health Insurance Markets, 2001: Structure and Change."

[204] Jean Sherman Chatzky, "Choosing Health Care," *Money*, 27 (1998): 152.

[205] Kevin J. Soo Hoo, 21.

[206] Kenny.

derived from two privacy notification studies administered in 2001. These studies were not specific to the health care industry but queried people's responses to offline privacy notices from financial institutions including banks, insurance companies, and credit card organizations. The studies explored people's response to such privacy notices. While the studies were not limited to insurance companies, the survey contexts and our contexts do overlap to a degree. One 2001 survey of 2468 adults indicated 25% (n=617) either "frequently read" or "always read" the privacy notices from such institutions (defined as institutions that send out a bill, credit card, bank, or other financial statement, as, for example, a health insurer, which deals with the financial aspects of healthcare, may).[208] A second 2001 survey of 2053 adults indicated that approximately 12% (n=246) "most of the time…carefully read" the privacy notices concerning their financial institutions (defined as banks, investment companies, or insurance companies).[209] This second survey also queried all respondents, asking what action they would take after reading a privacy notice sent to them from their financial institution. Nineteen percent would use "more discretion in choosing which financial institution with which to interact."[210] We can combine the results of both surveys to obtain an estimate in our model of the percentage of individuals who might "frequently" or "always" read such privacy notices and who might react based on the content of such notices. The combined estimate of the percentage of people who would "frequently" or "always" read such notices becomes 19%. If we assume that the 19% of all individuals who use more discretion in choosing a financial institution after reading a notice, from the second survey, is distributed uniformly across the different categories of readership regarding privacy notices within the second survey, then our hypothetical model would indicate that approximately 3.6% of adults would read a privacy notice and react based on content.

What is the impact of the roughly 3.6%? This is the number of people who would switch to the insurer that offered extra BQMA privacy protection. We focus on the larger insurer first but then focus on the smaller insurer. Imagine that the larger insurer installs additional BQMA privacy safeguards before the smaller one. It wants to explore any potential benefits of offering more policy-holder privacy. During annual reenrollment, about 3.6% of current enrollees of the smaller insurer would notice the larger insurer's privacy practices and switch health plans, if the smaller insurer lacked such practices.

How would policy-holders notice the larger insurer's practices? They may read about them in the insurer's information distributions. In the interests of obtaining privacy benefits the larger insurer may want to advertise its practices. One advertising method

[207] Joan Feigenbaum, Michael J. Freedman, Tomas Sander, and Adam Shostack, "Economic Barriers to the Deployment of Existing Privacy Technologies," 2002, 2, <http://citeseer.ist.psu.edu/cache/papers/cs/26515/http:zSzzSzcs-www.cs.yale.eduzSzhomeszSzjfzSzPrivacy-Barriers-WP.pdf/feigenbaum02economic.pdf> (22 October 2005).

[208] Mary J. Culnan and George Milne, "The Culnan-Milne Survey on Consumer & Online Privacy Notices," December 2001, <http://www.ftc.gov/bcp/workshops/glb/supporting/culnan-milne.pdf> (1 April 2005).

[209] Harris Interactive, "Privacy Notices Research: Final Results," December 2001, <http://www.ftc.gov/bcp/workshops/glb/supporting/harris%20results.pdf> (3 April 2005).

[210] Harris Interactive, "Privacy Notices Research: Final Results."

would be through workplace marketing efforts. These are informational efforts wherein the insurer relays its health benefit offerings through employers and allows employees to select more individualized coverage if they wish.[211] The insurer may also publicize its privacy practices in its Notice. HIPAA requires covered entities to send Notices to their constituents as per HIPAA's tenet 3, described earlier. Roughly 3.6% of individuals are attuned to and may take action based on such "notices." About 3.6% of policy-holders would notice the differences between the two insurers based on the content of any larger and smaller insurer's information distributions and switch health plans to the preferable insurer.

2.2.1.1 Plan Switching Complexity

We should mention that health plan switching may be more complicated than as explained above. Health plan benefits and price as well as privacy practices may be reasons why people switch plans. Further, how well privacy protection is advertised; how quickly it's implemented; and if it's provided for other internal insurer applications, which would enhance privacy protection, would all need to be computed to understand plan switching behavior.[212] Still, the roughly 3.6% computed above is triangulated by several evidentiary sources. In the 1999 California HealthCare Foundation Survey, a similar number of people did not merely *say* they would act based on PHI protection practices for applications such as the BQMA; they actually acted accordingly, in that case paying out of pocket to protect privacy in such applications. We will examine these data below. In a 2000 survey, 5% of the public used special software that hid its identity from the websites it visited.[213] Several health industry organizations have pointed out recently that patients may be concerned about disease management's privacy implications, and 1%-2% of individuals have opted out of such programs at one large health insurance organization recently potentially due to privacy.[214] [215] A lawsuit was filed in 2003 against the US Department of Health and Human Services alleging that the removal of the consent requirement for Treatment, Payment, and Operations in the latest version of HIPAA was a privacy violation.[216] The plaintiff, representing individuals and health care practitioners, wanted more control over PHI within health institutions. The plaintiff was a coalition of consumers and health care practitioners representing approximately 750,000 individuals in the US.[217] According to one of the lead plaintiff attorneys, the number of individuals

[211] See for example Allstate, Workplace Division, "Workplace Marketing," <http://www.ahlcorp.com/ProdIndWork.asp> (Apr 1, 2005).
[212] For example, see analogous concepts in "Living Large," *Health Management Technology*, 24 (2003): 32-33.
[213] The Pew Internet & American Life Project, 10.
[214] Laura Benko, "Long-range Forecast: Partly Healthy, Chance of Storms," *Modern Healthcare*, 34 (2004): 28.
[215] William Atkinson, "Making Disease Management Work," *Society for Human Resource Management*, 47 (2002), <http://www.shrm.org/hrmagazine/articles/0102/0102atkinson.asp> (24 July 2005).
[216] Deborah Peel, "Lawsuit Challenges HIPAA," *MSPP News*, 13 (2003), <http://www.mspp.net/hipaa_lawsuit.htm> (30 August 2005).
[217] Peel, "Lawsuit Challenges HIPAA."

who might have been represented could have been several million as the coalition had to turn away organizations to efficiently coordinate the litigation.[218] If up to several million individuals were concerned, up to one percent of the US population in 2003 were concerned.[219]

Profitability can be computed as follows. The larger insurer would capture 0.036 * 268,000 or about 9640 people from the smaller insurer during annual re-enrollment.[220] In 2003, the typical profit per member per month for an insurer of any size was $5.15.[221] The annual profit therefore becomes $61.8 per member. The annual profit to the larger insurer from attracting such individuals becomes 9640 * 61.8 or approximately $595,000 in 2003. This number must be converted to 2001 figures for consistency. Using the Consumer Price Index, the total becomes about $572,000.[222] The smaller insurer would lose approximately this same sum. It would continue losing such a sum annually until it also offered similar privacy protections. Of course, if the smaller insurer installed such protections first, its potential gains (and the corresponding losses of the larger insurer) would be considerably larger. The same percent of people from the larger insurer would now switch to the smaller insurer. However, such gains would have to be moderated by the smaller insurer's capacity. The smaller insurer may not be able to take on 0.036 * 748,000 or about 26,900 additional policy-holders quickly due to infrastructure limitations. Nevertheless, a considerably larger profitability and loss for the smaller and larger insurers, respectively, is possible if the smaller insurer installs such protections first. We examine the probability of such profitabilities later in the text.

2.2.2 Reduction in Loss from Information Abuse

Next we measure the loss from having an employee abuse identifiable PHI because he has access to it. This quantifies the intangible "increase in operating costs" an organization may face if it offers less privacy protection, one loss of not providing strong protection.[223] [224] Preventing the loss would create the benefit. Data on the precise magnitude of such losses are difficult to find. However, an estimate can be derived for the purposes of our model using a combination of an unpublished data set of privacy violations at East Coast health institutions, spanning over 15 years, made especially

[218] Jim Pyles, attorney, telephone interview with author, July 26, 2005.

[219] US Census Bureau, "National and State Population Estimates," <http://www.census.gov/popest/states/tables/NST-EST2005-01.xls> (Microsoft Excel file, 28 January 2006).

[220] Once again, as per a reference earlier, it is assumed that children belonging to the same health plan as their parents would go along with the health plan choice of their parents, who would be making the choice to switch during annual re-enrollment.

[221] National Association of Insurance Commissioners finance staff, email to author, April 7, 2005.

[222] Federal Reserve Bank of Minneapolis, "What is a Dollar Worth?" <http://minneapolisfed.org/research/data/us/calc/> (13 April 2005).

[223] G. Stevenson Smith, 47.

[224] Thomas R. Shaw, "The Moral Intensity of Privacy: An Empirical Study of Webmasters' Attitudes," *Journal of Business Ethics*, 46 (2003): 307.

available for this thesis research, and a public survey with less perfect data. The data set regarding East Coast health institutions will be called the East Coast malpractice data set from now on due to its malpractice costs information content.

The East Coast malpractice data set contains data on the legal and administrative costs to manage the breach of confidentiality claim and any damages paid out for the East Coast health institutions.[225] Other intangible losses, such as a particular institution's lost market share or decline in reputation, are not in the data. I performed a content analysis of the violations in the East Coast malpractice data set and included incidents in which an employee obtained PHI in paper or electronic form, as opposed to obtaining PHI by, for example, treating patients if he was a clinician. I also included cases where an employee used PHI without consumer authorization or in an obviously abusive manner, such as stealing someone's identity, as opposed to using PHI under a legal context where consumer "rights" might be less enforceable. In this case, the consumer might not have privacy protection "rights" even if better protection were to be implemented. For each year in the data set, I spread the costs across all employees, across all institutions. For a given year, the loss per employee was never more than one or two dollars.[226] For a number of years it was considerably smaller.

The second data source, the 2001 CSI/FBI survey, enabled us to come up with an estimate of per employee losses related to technical and infrastructural patches required after an insider security breach. The data from this survey only appear to focus on the financial costs to recover from breaches at the technical infrastructural level and do not quantify intangible losses.[227] We extract the information necessary from the CSI/FBI survey to compute the per-employee losses. The CSI/FBI survey lists 10 categories of IT violations into which respondents may place their annual financial losses.[228] Some of these categories, such as Sabotage or System Penetration, were mentioned before. Although there is not an exact one-to-one mapping between any of these specific

[225] Unpublished data from Executive Information System database, Controlled Risk Insurance Company (CRICO)/Risk Management Foundation (RMF). Obtained on March 4, 2005.

[226] Note, the actual computation was as follows. We need a loss per employee per year. We must handle the case when organizational employees can only cause damage to their own, not other organizations. Risk must be properly apportioned to an organization's own employees. Thus an "average organization" was created. It contained an average number of employees and an average IT abuse loss per year. We use the computations associated with the "average organization" to obtain an average loss per employee in the East Coast malpractice data set. We first found the average number of employees for the average organization. We divided the total number of employees across all organizations by the total number of organizations in the data set. We next obtained "average loss" for the average organization. We divided the total losses across all organizations for every year by the number of organizations in the data set. Finally, to obtain the average loss per employee in the data set, the average organization's average loss was divided by its average number of employees. However, the number of organizations in the data set canceled each other out in these computations. The average number of employees and average loss were both divided by the total number of organizations: since the average loss is divided by the average number of employees, the total number of organizations is divided by itself, leading to 1. The final computation in the text representing average IT abuse loss per employee can therefore be simplified to: total losses across all organizations per year divided by all employees across all organizations, as presented in the text.

[227] Computer Security Institute, "2005 CSI/FBI Computer Crime and Security Survey," <http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2005.pdf> (30 August 2005).

[228] See Computer Security Institute, "2001 CSI/FBI Computer Crime and Security Survey."

categories and IT losses due to PHI access we will use the "Financial Fraud" category for our analysis. Another 2004 e-crimes survey regarding PHI abuse appears to place internal PHI abuses into a similar category. Internal employees could abuse PHI in several ways, such as committing identity fraud or improperly disposing of PHI.[229] [230] [231] The authors of the 2004 E-Crime Watch Survey placed a number of such incidents into a "fraud" category of their own for analysis.[232] We mimic this approach, utilizing "Financial Fraud" within the CSI/FBI survey. Of the CSI/FBI respondents who reported a monetary loss in the Financial Fraud category, the average loss per organization was $4,420,738.[233] Further, of all breaches, approximately 50% were due to *insider* attacks.[234] Therefore, the average loss per organization from insider loss was roughly $2,210,000.

We must also know the number of employees per respondent to quantify per-employee losses. The CSI/FBI survey collected data on the 534 company respondents. The number of employees varied widely.[235]

Percent of the 534 CSI/FBI company respondents	Number of employees per respondent
16%	1 - 99
16%	100 - 499
8%	500 - 999
22%	1000 - 5000
11%	5001 – 9999
27%	10,000 or more

Company respondents and their number of employees

For the purposes of our hypothetical model, let's assume the number of employees was distributed normally within these ranges and we'll take the mean number of employees for each of these ranges. Thus: 16% of the respondents had 50 employees; 16% had 300; 8% had 750; 22% had 3,000; 11% had 7,500; and 27% had 10,000 or more employees. To obtain a conservative upper limit on the number of employees in companies with 10,000+ employees, I will use the 2001 data from the list of the largest Fortune 500 companies: the top 10 companies had on average 436,300 employees, assuming a normal distribution of employees.[236] Therefore, for our model, the average number of employees per organization becomes about 119,000. Using the previously derived estimate of organizational losses due to insider breaches, and this estimate of the average number of employees per organization, we can now estimate the annual technical/infrastructure loss per employee due to internal attacks to be 2,210,000 / 119,000 or about $18. Combining this technical infrastructure loss with the breach of confidentiality management costs

[229] HIPAAps Privacy and Security.

[230] CRICO/RMF data.

[231] CSO Magazine, "2004 E-Crime Watch Survey," <http://www.cert.org/archive/pdf/2004eCrimeWatchSummary.pdf> (10 April 2005).

[232] CSO Magazine.

[233] Computer Security Institute, "2001 CSI/FBI Computer Crime and Security Survey."

[234] Computer Security Institute, "2001 CSI/FBI Computer Crime and Security Survey."

[235] Computer Security Institute, "2001 CSI/FBI Computer Crime and Security Survey."

[236] "Fortune 500 Largest US Corporations," *Fortune*, F-31, April 15, 2002.

derived from the East Coast malpractice data set we come up with a global estimate of per-employee loss due to identifiable PHI access. This estimate is no more than $20 per year.

The following are the cost implications. From one PM vendor, the number of people required to operate PM may be from 2 to 5.[237] The larger values would hold especially for larger organizations. Such people run the PM platform, ensure data integrity, and perform other functions. If an organization develops and programs its own PM application, an additional approximately 3-4 people annually may be required to provide the needed business, clinical, and IT expertise.[238] We use 4 people for our computations; this includes the average number of basic staff; this also includes less than one in-house PM platform development staff person. Assuming operational similarity across the BQMA, as explained in Section 1.3.2, for the larger insurer, a total of 16 people might be employed operating the four BQMA. The upper bound on the annual loss due to insider PHI abuse becomes roughly 16 * $20 or $320. This will be approximately true for the smaller insurer, too, although potentially lower.

2.2.3 Paying Out-of-pocket Dynamics

The next intangible loss we quantify represents the impact of using privacy-protective policies and technologies on an organization's reputation. If an insurer implements procedures to better protect privacy in the hopes of improving its reputation, it may reverse people's "defensive" behavior. People are paying out of pocket to protect privacy. If the insurer offers better privacy protection, it may reverse people's payment behavior, increasing the insurer's cost as the insurer must pay for policy-holders' healthcare. The insurer can target different groups of policy-holders in attempts to recover the funds necessary to pay for the extra claims expenses.

Consider people's "willingness to pay" (WTP) for BQMA anonymity. One of the California HealthCare Foundation Survey questions specifically asked respondents whether they paid out of pocket instead of submitting claim records to their insurer to avoid having their "employer or someone else" gaining access to their medical information. Looking at these responses as a function of the type of insurance a respondent had, 11.4% of individuals (61 out of 535 individuals) who were in "strict" managed care paid out of pocket to prevent such a disclosure; approximately 5.1% of individuals (i.e., 55 out of 1073 individuals) who were in "loose" managed care paid out of pocket to prevent this disclosure; and 4.2% of individuals (9 out of 217 individuals) who were in "traditional

[237] MEDecision staff, telephone interview with author, March 22, 2005.

[238] Some organizations rely on such in-house development although the trend may not be rising. (Director, Clinical Informatics, large health insurance organization in the South East, telephone interview with author, April 1, 2005). As the PM market matures, vendors are improving PM platforms, requiring less need for organizations to develop their own applications. Still, if the insurer wants to differentiate itself in the health insurance market or serve special populations it might modify the PM application in-house.

plans" paid out of pocket to prevent such a disclosure.[239] We will combine the "loose" managed care and "traditional plans" into a new derived category. The percent of individuals in the "looser" managed care arrangements is approximately 4.96%. The difference between the "looser" managed care arrangement and "strict" managed care is 6.44%. Why this more than double difference in out-of-pocket payment? One possible explanation is BQMA information practices which vary between these two insurance groups. As some analysts point out, managed care organizations rely on applications such as the BQMA to contain costs.[240] [241] For example, Utilization Review can be used to approve a referral by a primary care physician or approve treatment for patients, as explained before. The California HealthCare Foundation Survey defines individuals to be in "strict" managed care if their health insurer enforces similar practices. "Strict" managed care individuals are those who pay less "if [they go to] a doctor from a list, but...pay more if [they] go to a doctor not on the list... [and they] sign up with a specific primary care doctor or group of doctors who provide all [their] routine health care...*[and they must] have a referral by a primary care doctor before [they] can see a medical specialist...[or they must] have approval or a referral before [the health plan] will pay for any of [their] costs* for visiting a doctor who is not in the plan."[242] That is, those in strict managed care may be aware of the greater oversight provided by their insurance company using applications such as Utilization Review and this may be one of the explanations of the different responses regarding privacy protection within the California HealthCare Foundation Survey.[243] [244]

We assume that of people in strict managed care who are paying out of pocket, 1 - 0.0496, or approximately 95% will pay out of pocket due to BQMA information practices. We remove from the strict managed care group all out-of-pocket payment behavior regarding those in "non-strict" managed care arrangements. That is, we remove those in "looser" managed care arrangements who presumably would not be worried as much about BQMA information practices. The total percent of individuals in strict managed care paying out of pocket only due to the BQMA information practices becomes 0.95 * 0.114 or approximately 10.8%.

We quantify the WTP for analysis. A secondary analysis of the California HealthCare Foundation Survey revealed that in 1998 of the respondents who said they paid out of

[239] Larry Hugick, staff, Princeton Survey Research Associates International, fax to author, December 11, 2003.

[240] See Kremer, 553-554.

[241] See Bradford Kirkman-Liff, "Restoring Trust to Managed Care, Part 1: A Focus on Patients," <http://www.ajmc.com/files/articlefiles/AJMC2003feb1Kirkman174-180.pdf> (1 April 2005).

[242] "Strict" managed care is defined as answering affirmatively to questions 48 and 49 as well as to questions 50 or 51 in the California Healthcare Foundation Survey. (Larry Hugick, staff, Princeton Survey Research Associates International, telephone interview with author, March 21, 2006; California Healthcare Foundation, "Medical Privacy and Confidentiality Survey," 1999, <http://www.chcf.org/documents/ihealth/topline.pdf> (Apr 1, 2005)). The verbiage for these questions in the Survey is in the text.

[243] See, for example, Kremer, 553-554, 556.

[244] See Kirkman-Liff, "Restoring Trust to Managed Care, Part 1: A Focus on Patients."

pocket and were in strict managed care, the following sums were paid in order to avoid divulging their PHI to an "employer or someone else:"[245]

—63.2% of individuals paid 0
—15.2% of individuals paid less than $100
—10.1% of individuals paid $100 - $500
—4.2% of individuals paid $500 - $1000
—3.6% of individuals paid $1000 - $5000
—3.8% of individuals refused to answer

Assuming that the dollar amounts were normally distributed within each range and taking the mean within each of these ranges, we come up with the number of dollars paid out of pocket by each of these respondents—the total is about $170. We conservatively assume the same behavior in 2001 for such individuals because privacy concerns have not abated and might have even increased since 1998, the year of the California HealthCare Foundation Survey. Converting via the Consumer Price Index, the 2001 value is about $180. The California HealthCare Foundation Survey also indicates that 28% of all insured individuals nationwide belonged to strict managed care in late 1998. Therefore, we assume 28% of the 10.8% or about 3.02% of individuals nationwide were concerned about the BQMA privacy practices regarding an insurance organization.[246] We assume membership in strict managed care should be the same in 2001 as in late 1998, only two and a half years later.[247] We return to our model of the hypothetical larger insurer. It's approximately 750,000 policy-holders annually pay 748,000*180*(0.0302) or about $4,060,000 out of pocket to avoid identifiable PHI use for the BQMA.[248] The policy-holders at the smaller insurer have a proportionally smaller WTP.

[245] Larry Hugick, fax to author. In fact, the secondary analysis provides these statistics for individuals *and* their immediate families paying out of pocket. Since we are computing for individuals only, I conducted another secondary analysis of the data. (Princeton Survey Research Associates staff, email to author, June 20, 2003). The percentages for individuals paying out of pocket across the payment categories in the text, such as paid less than $100, paid $100-$500, etc., as will be shown in the text, were roughly similar to those of such categories representing individuals and their immediate family members. Therefore, we assume strict managed care statistics for individuals should be similar to strict managed care information for individuals and immediate family members. We rely on the secondary analysis statistics for individuals and immediate family members to present, in the text, individuals' payments out-of-pocket regarding strict managed care.

[246] Note, the assumption is that often health insurance organizations offer multiple products, such as "traditional" (indemnity) plans and various types of managed care offerings. (See James Robinson, "The Future of Managed Care Organization," *Health Affairs*, 18 (1999): 7-24). Therefore, for a given insurer, some policy-holders should purchase its "strict" managed care offerings, some its "loose" managed care offerings, etc. This allows us to use the 28%, which represents the percent of all insured individuals belonging to strict managed care, as implying what percent of a typical insurer's policy-holders purchase its strict managed care products. Afterwards, we can apply the 10.8%, which represents people's WTP behavior towards organizations which have only strict managed care products, to the 28% for a typical insurer, which has strict managed care as well as other products. The result is the percent of people nationally concerned about strict managed care BQMA information practices within an insurer.

[247] California Healthcare Foundation, "Medical Privacy and Confidentiality Survey," 25.

[248] We assume that the children who belong to the roughly 748,000 members of the large insurer also have a "willingness-to-pay" for privacy. That is, the assumption is that the children may pay for their care themselves, as implied by Ford. (See Carol Ford, Abigail English, and Garry Sigman. "Confidential Health Care for Adolescents: Position Paper of the Society of Adolescent Medicine," *Journal of Adolescent*

At first, such a WTP appears to benefit the insurer. The health plan avoids paying roughly $4 million of its own resources; it avoids this large loss.[249] [250] However, such a perception is reversed if the insurer installs and publicizes the new BQMA privacy protections to obtain any benefits from such protections. It might suddenly face an increase in annual claims of about $4 million. People who paid out of pocket to protect privacy may now stop paying out of pocket and start submitting claims. They are getting BQMA privacy protection, which is what they wanted, for free, by relying on the insurer's new advertised process.

2.2.3.1 Managing Out-of-pocket Payments

The significance of the $4 million to an insurer is unclear. The expense might be large. The financial literature points to the "medical loss ratio" as one indicator of an insurer's financial performance. This ratio is computed by dividing health care expenses, the claims paid out, by the total premiums collected by the insurer.[251] A rise in a few percent or even half a percent of the medical loss ratio might be meaningful to an insurer, potentially signifying that the insurer cannot control expenses.[252] The WTP might produce a small rise in the medical loss ratio. For example, for two large health insurers in New England, Harvard Pilgrim and Blue Cross and Blue Shield of Massachusetts, in 2004, a roughly $4.3 million in additional claims (2004 dollars) would signify a roughly 0.2% or 0.1% change in the medical loss ratio, respectively, which might be significant.[253] [254] In 2004, profits for Harvard Pilgrim would fall from $38,619,000 to about $34 million, a drop of about 11% because of such an expense.

On the other hand, the $4 million might be less of a concern to very profitable insurers. Several large health insurers had net incomes from over $240 million to over $340 million in the early 2000s and in 2004, of which the roughly $4 million would be a very small percent.[255] [256] If the $4 million impact is not financially significant, the insurer might absorb it.

Health, 35 (2004): 160-7). Alternatively, parents may pay for their children's care to protect the privacy of the children's medical information. Both assumptions are also made regarding the smaller insurer.

[249] Laura Benko, "Less is Not More," *Modern Healthcare*, 30 (2000): 41.

[250] Laura Benko, "...You Pay," *Modern Healthcare*, 33 (2003): 8.

[251] Paul Grimaldi, "The Versatile Medical Loss Ratio," *Nursing Management*, 29 (1998): 12.

[252] Joe Niedzielski, "Rising Expenses Nip at Results of Public HMOs," *National Underwriter*, 100 (1996): 4, 9.

[253] Harvard Pilgrim Health Care, "Annual Report 2004," <http://www.harvardpilgrim.org/hpimages/HP-2004Annual.pdf?SMSESSION=NO> (30 August 2005).

[254] Blue Cross and Blue Shield of Massachusetts, "Annual Report 2004," <http://www.bcbsma.com/common/en_US/repositories/CommonMainContent/aboutUs/AnnualReport/BCBSMA_04_Financials.pdf> (30 August 2005).

[255] Laura Benko, "Earnings at a Premium," *Modern Healthcare*, 32 (2002): 22-23.

[256] Blue Cross and Blue Shield of Massachusetts.

If the insurer wants to recover the new dollars lost due to increased claims from the population who were previous paying out of pocket, there may be several recovery methods. The easiest would be to divide the $4 million by all policy-holders and raise everyone's premium by the same amount. Everyone would pay for privacy protection desired by some. In 2001, an average covered employee paid $360 per year for health insurance.[257] In 2001, adults represented 69.1% of all the non-elderly individuals who had health insurance.[258] We focus on adults since they, as opposed to children, would probably be the ones paying for health insurance, including any for that of their families. The large insurer therefore had 748,000 * 0.691 or about 516,000 adult members. If the WTP were fully passed on to such employees, 4,060,000 / 516,000, or about $7.86, would be the additional annual cost to such employees, policy-holders of the large insurer. Note, this assumes the employee is responsible for the entire premium cost. If the *employer* pays for part of the premium, as is common, the employer would be responsible for part of the premium increase.[259] Compared to the $360, the $7.86 would represent an annual payment increase of approximately 2.1% for health insurance for the adults.[260]

The insurer can target groups that contain individuals with a WTP and charge them premiums to cover the WTP. The author named *Perry 6* lays out a psychological framework of how individuals may react to risk.[261] Some look at risk fatalistically, others try to encourage the passing of legislation to protect against risk, and still others protest risks. *Perry 6* segregates individuals' reaction to privacy concerns based on such risk profiles. Some perceive privacy exposures as demeaning, others look at data collection as a nuisance rather than a threat, and still others look at privacy risks as injustice or a violation of *principles*. The insurer can assess which of current employer groups are more likely to contain individuals with such risk profiles and charge them premiums incorporating their constituents' WTP. The insurer can identify other health coverage or services individuals would want based on the risk profiles. It can create new health benefit products offering such services in which premiums include the costs of the services as well as the costs of privacy protection. The insurer can collect the WTP as people join the new groups to obtain new health benefits.

The insurer can also try individualized approaches. Higher deductibles could be set to incorporate the WTP plus the regular deductible associated with health plan benefits in some of the more "individualized" insurance products that insurance organizations are

[257] The Henry J. Kaiser Family Foundation. "Employer Health Benefits, 2001 Annual Survey," 1.

[258] The Henry J. Kaiser Family Foundation. "Health Insurance Coverage in America, 2001 Data Update," 21.

[259] For example, see the Henry J. Kaiser Family Foundation. "Employer Health Benefits, 2001 Annual Survey," 1.

[260] In fact, this computed 2.1% would probably be smaller. The $360 represents individual coverage. If family coverage would be taken into account, representing adults paying for their children, the annual premium would be considerably higher. (See the Henry J. Kaiser Family Foundation. "Employer Health Benefits, 2001 Annual Survey," 1). Thus, the computed $7.86 would represent considerably less than one percent of this higher premium, and thus an even smaller relative payment, for each adult.

[261] Perry 6, "Who Wants Privacy Protection and What Do They Want?" *Journal of Consumer Behavior*, 2 (2002): 86-87.

considering today.[262] [263] Under such coverage, employees can choose more tailored health benefits rather than the more homogenous products typically offered via managed care. When a consumer first sees the provider, she can pay out of pocket for the deductible, into which her WTP would be incorporated.

2.2.3.2 Policy-Holder Welfare Maintained

Note, when individuals or small groups are targeted with increasing expenses to incorporate the WTP, the targeted parties should not fear the increased expense. The insurer should publicize the equivalent if not better insurance environment the policy-holders may be gaining should they stop paying out of pocket and start submitting claims. Financially, policy-holders should not be worse off. If, in the past, they paid amount X for premiums and Y out of pocket to protect privacy, now they will be charged a premium not far from X+Y, a similar outlay. However, the additional benefit policy-holders can experience is better insurance services. By allowing the insurer to know information about policy-holders, because they are now submitting personal health information, the insurer may create better health benefits for them. Recall one goal of the insurer may be improved policy-holder care quality. The insurer can better learn about its policy-holder base and its needs to create better insurance products. It will certainly want to do so in a competitive environment, wherein through better services it may attract more customers. Indeed, one evident benefit to policy-holders from submitting claims data is the more optimal administration of disease management which can improve policy-holder care quality, as will be discussed below.

The insurer will most likely acquire sufficient funds to pay for all newly incoming claims from policy-holders under the above solicitation methods, and may capture the consumers' WTP. The smaller insurer will face considerably less WTP withholding as it has fewer policy-holders.

2.2.4 Disease Management Implications

The insurer will need to encourage the submission of claims to itself as otherwise it may suffer a different financial loss. Using the same WTP analysis as above, we quantify a different intangible "increase in operating costs" to the organization. Out-of-pocket payments may undermine the creation of records needed by a PM-like process, lessening the financial savings from the associated disease management program to the insurer. We note, however, the loss computed below will not be large, but will be buttressed by other losses in the discussions later on.

[262] Stephen Parente, Roger Feldman, and Jon B. Christianson, "Employee Choice of Consumer-driven Health Insurance in a Multiplan, Multiproduct Setting," *Health Services Research*, 39 (2004): 1092, 1095, 1106-1107.

[263] James Robinson, "Reinvention of Health Insurance in the Consumer Era," *JAMA*, 291 (2004): 1880.

We examine one possible disease management-like program as a case study: using telemedicine to manage women with high-risk pregnancies. Using telemedicine can significantly reduce the costs of managing high-risk pregnant women to the insurer. With the use of a special device attached to the woman's abdomen to electronically monitor her growing fetus, and subsequent transmission of these data through a telecommunication line to a nurse assessing her symptoms, a telemedicine program can identify pregnancy abnormalities when the woman is at home far from a care provider. Appropriate interventions can be initiated based on symptoms by staff monitoring her care. Women can be selected into a telemedicine program by a PM-like process that can monitor their claims or other data signifying their high-risk status. Those identified as high-risk can be enrolled in telemedicine. Women have privacy concerns regarding their high-risk pregnancies. Another analysis of the California HealthCare Foundation Survey shows that for people who were ill in some way, there was a slightly higher likelihood that they paid out of pocket to prevent an "employer or someone else" from knowing their PHI.[264] A 2002 study of prenatal practices in New York State showed that of women in prenatal care, those with the highest education attained, women of color, women who were older, or those who presented late for prenatal care were all more likely to not want to share birth information of their children with their obstetrician, pediatrician, and particularly the New York State immunization registry.[265] The women's reasons could reflect suspicion of government use of PHI or concern about PHI use for "unknown" reasons. The California HealthCare Foundation Survey suggests women will pay out of pocket to protect their privacy, as explained in Section 2.2.3. PM will not be able to locate the data needed to enroll the high-risk pregnant women into disease management because complete data are not available to the application. We examine this phenomenon using a particular telemedicine study.[266]

We should mention the key assumptions we make in the analysis below. We first assume that the PM platform used to identify the high-risk pregnant women relies on risk assessment data for identification. Risk assessments are questionnaires or associated pregnancy tests identifying the risk factors that might place women at risk for poor childbirth. If there are no risk assessment visits, i.e., all the women's prenatal care visits collect roughly the same or non-specialized data, or PM relies on all prenatal data for identifying such women instead of relying on only the specialized risk assessment data, the insurer might still get the financial benefits of disease management despite women's defensive behavior regarding disease management. PM will rely on other "identification"

[264] Larry Hugick, fax to author.

[265] Timothy Dye, Martha Wojtowycz, Mary Applegate, and Richard Aubry, "Women's Willingness to Share Information and Participation in Prenatal Care Systems," *American Journal of Epidemiology*, 156 (2002): 288.

[266] We should point out that although we are using telemedicine as an example for this analysis it is not necessarily the best current practice and may not always be currently administered for women with high risk pregnancies. However, interventions for high risk pregnant women with similar cost implications as telemedicine are being utilized. Therefore, our analysis, although focused on telemedicine is still relevant because even under other interventions, cost implications to the insurer, as will be described in the text, may be the same. (Marianne Hutti, Doctor of Nursing Science, Professor, School of Nursing, University of Louisville, email to author on November 15, 2005).

data, e.g. data from other prenatal care visits, to identify the women and enroll them in telemedicine despite their privacy-motivated out-of-pocket payments, creating financial benefits from disease management for the insurer.

A related assumption is that each risk assessment visit will generate one record which will feed the PM software. One author demonstrates the atomicity of electronic risk assessment data.[267] Another author implies that risk assessment data are collected on a single paper form.[268] Therefore, we might assume one transcribed electronic form might be consequently generated. If there is more than one record for each risk assessment visit, then from an identifier error point of view, as will be shown below, the non-linkage of some records due to record identifier errors may not completely prevent a high-risk woman from being identified for telemedicine. The insurer will therefore obtain some financial benefits from the existing context, which would not pressure the insurer to offer additional privacy protections by reducing errors. Other data might suggest a woman will not have a "normal" pregnancy because multiple records are available for each risk assessment visit. PM might identify the woman to some degree because these records can be linked, potentially creating financial benefits from disease management for the insurer despite any errors in linkage identifiers.

Another assumption is that women can recognize the difference between risk assessment visits and other prenatal care. For example, they may recognize the different pregnancy tests they must complete or the different questions from providers they must answer during the risk assessment visits. If the women cannot separate the risk assessment visits from other prenatal care, the risk assessment data may remain intact because the women may be paying out-of-pocket for *different* visits. The insurer would again feel less financial pressure to add privacy protection to the BQMA because it is not losing money in the current context. PM can enroll the women in disease management because needed "risk" data are not absent since women cannot differentiate the risk assessments visits from other prenatal care. Therefore, PM again can enroll the women in telemedicine and create financial benefits from disease management for the insurer because the women are paying out-of-pocket for non-risk assessment visits.

Finally, the women also want to pay *for* the risk assessment visits as opposed to paying for as many initial prenatal care visits as they can to protect their privacy. The California HealthCare Foundation Survey states that medical visits are paid for out of pocket due to privacy concerns but it doesn't specify which visits are paid for out of pocket. If the women chose to pay for all initial prenatal care visits to protect their privacy by preventing the transmission of their PHI from the beginning of their pregnancy they might not be able to afford to pay for the risk assessment visits. The $180 out-of-pocket payment computed using the California HealthCare Foundation Survey in Section 2.2.3

[267] Michael Ross, Catherine A. Downey, Rose Bemis-Heys, Men Nguyen, Debbie L. Jacques, and Gary Stanziano, "Prediction by Maternal Risk Factors of Neonatal Intensive Care Admissions: Evaluation of >59,000 Women in National Managed Care Programs," *American Journal of Obstetrics and Gynecology*, 181 (1999): 835.

[268] John Morrison, Niki K. Bergauer, Debbie Jacques, Suzanne K. Coleman, and Gary J. Stanziano, "Telemedicine: Cost-Effective Management of High-Risk Pregnancy," *Managed Care*, 10 (2001): 43-44.

would only cover several initial prenatal visits. By the time the subsequent risk assessment visits arrive the women may no longer have sufficient funds. Again, there would be less financial pressure on the insurer to add privacy protection to the BQMA because it is not losing money in the current context. PM should find the risk assessment data and enroll the women in telemedicine because they cannot afford to pay out-of-pocket for the risk assessment visits, again creating financial benefits from disease management for the insurer because risk assessment data are present.

2.2.4.1 Telemedicine and Privacy Interactions

In 2001, Morrison et al. retrospectively analyzed data demonstrating the cost savings to an HMO when the HMO offered telemedicine services to high-risk pregnant women.[269] This study, called "2001 study" from now on, examined 1992-1994 health outcomes and cost data to understand one HMO's efforts to reduce preterm births. Based on a diagnosis of preterm labor, the HMO identified women with preterm labor and suggested that a telemedicine program be adopted for their care. This program linked a woman's home to a care provider and transferred her home uterine activity monitoring, via a device attached to her abdomen, over a standard telephone line to a patient service center. The information was interpreted by an obstetric nurse who assessed the woman's symptoms of preterm labor. The 2001 study compared clinical and cost outcomes for two similar high-risk groups: women who received telemedicine services and those who did not. For the control group the average cost of the pregnancy was $21,684, while for the intervention (telemedicine) group the average cost was $7225.

Although the 2001 study used the diagnosis of preterm labor as the method to identify the women, enrollment into telemedicine also happens through risk assessments.[270] [271] Poor risk "scores" on such questionnaires or on corresponding pregnancy tests imply higher probability of pregnancy complications. Indeed, in the 2001 study, the HMO analyzed data collected before the preterm labor diagnosis and prescribed telemedicine if appropriate, suggesting such data can be useful for needed intervention.[272] There are two, sometimes three, risk assessments done for all pregnant women to understand any risks associated with their pregnancies.[273] [274] [275] [276] [277] We will use two assessments for our

[269] Morrison, 42-49.

[270] Morrison, 43.

[271] Michael Corwin, Susan M. Mou, Shirazali G. Sunderji, Stanley Gall, Helen How, Vinu Patel, and Mark Gray, "Obstetrics: Multicenter Randomized Clinical Trial of Home Uterine Activity Monitoring: Pregnancy Outcomes for All Women Randomized," *American Journal of Obstetrics and Gynecology*, 175 (1996): 1281.

[272] Matria healthcare staff, telephone interview with author, March 24, 2005.

[273] Ross, 836.

[274] Deanna Lear, Laura C. Schall, Gary M. Marsh, Ken S. Liu, and Yvonne Yao, "Identification and Case Management in an HMO of Patients at Risk of Preterm Labor," *The American Journal of Managed Care*, 4 (1998): 866.

[275] Morrison, 43.

[276] Marianne Hutti and Wayne M. Usui, "Nursing Telephonic Case Management and Pregnancy Outcomes of Mothers and Infants," *Lippincott's Case Management*, 9 (2004): 290.

analysis. One risk assessment typically takes place before any home uterine activity monitoring is administered; the other at approximately the same time as such monitoring is administered. Typically, home uterine monitoring is administered at 24 weeks gestation.[278] In the 2001 study, the HMO performed three risk assessments during the pregnancy at 12, 24, and 30 weeks gestation to improve the management of pregnant women.[279]

A PM process can be set up to wait for the risk-assessment data. The data can be stored electronically.[280] [281] A PM or similar process can be set up to regularly monitor the digital data and find the collation of records that identifies women as high risk. The risk assessment data can indicate the degree of risk. The PM-like process can be run frequently so that identification may happen quickly, such as weekly or even daily, after which the women could be enrolled in a telemedicine program.[282]

Due to women's privacy concerns, the insurer may not record risk-assessment data for the first risk assessment and at times the second risk assessment. Women will either pay out of pocket for risk assessment visits or they may avoid their provider altogether during those prenatal care visits, limiting the transmission of risk assessment data to the insurer and thus to its PM platform. Risk assessments can be done by providers or non-providers.[283] [284] [285] Clinicians may be more capable of adding more detailed risk-assessment information, especially any clinically-related data.[286] At other times, the insurer's staff or disease management staff hired by the insurer to manage high-risk

[277] John J. Fangman, Peter M. Mark, Leslie Pratt, Kathleen K. Conway, Margaret L. Healey, John W. Oswald, and Donald L. Uden, "Prematurity Prevention Programs: An Analysis of Successes and Failures," *American Journal of Obstetrics and Gynecology*, 170 (1994): 744.

[278] Allison Kempe, Benjamin P. Sachs, Hope Ricciotti, Arthur M. Sobol, and Paul H. Wise "Home Uterine Activity Monitoring in the Prevention of the Very Low Birth Weight," *Public Health Reports*, 112 (1997): 433.

[279] Morrison, 43.

[280] Hutti, 290.

[281] See Ross, 835.

[282] Indeed, such a process may even be called "medical management" rather than Predictive Modeling per se. In this context, it's not uncommon for a provider to electronically request permission from the insurer, for example, to provide a certain medical service to a patient. The insurer's software can automatically grant permission based on the nature of the request and service arrangements between provider and insurer. (Implied, MEDecision staff, telephone interview with author, April 11, 2005). Such software can also be used for disease management-like monitoring and associated intervention. (See MEDecision, "Advanced Medical Management," <http://www.medecision.com/page.cfm?page=advanced> (12 April 2005)). However, even in today's PM context, health insurers are trying to devise PM-like systems as suggested in the text, such as, for example, Humana. (Matria healthcare staff, telephone interview with author, March 24, 2005).

[283] Health Alliance Plan. "Healthy Living, Prenatal Care Chart," <http://www.hap.org/healthy_living/teenadult/prenatca.php#High%20Risk> (17 April 2005).

[284] Lear, 866.

[285] Morrison, 44.

[286] Lear, 866.

pregnant women can perform the risk assessments, such as, for example, via phone.[287] [288] [289] [290] [291]

2.2.4.2 Paying for Risk Assessments, and Other Privacy "Defensive" Behavior

If the risk assessments are performed in provider offices, the California HealthCare Foundation Survey data imply that the women will pay for the risk assessments out of pocket, via the $180 WTP as computed before, to protect their privacy. The risk assessments suggest they have a high-risk pregnancy because the data collected may indicate the women's high-risk status. The women do not want a stigma. The California HealthCare Foundation Survey data suggest such women will pay the WTP sum but then start submitting claims forms to the insurer to suppress any information which might suggest their status.

The women can afford to pay for the risk assessments. The women's WTP covers the cost of more than one risk assessment visit. According to one source, the risk assessment visits may cost up to approximately 25% more than a regular prenatal visit as billed to the insurance organization. The physicians' offices must collect the extra risk factor data from the women.[292] We compute the cost of a regular prenatal visit to compute the cost of a risk assessment visit. According to the American College of Obstetrics and Gynecology, one visit every four weeks is recommended until 28 weeks gestation.[293] [294] One visit is recommended every two weeks from 28 through 36 weeks gestation; and one weekly visit after that. On average, in the 2001 study, the pregnancy itself across both control and intervention groups lasted about 36.8 weeks. Therefore, for the first 28 weeks there should have been 7 prenatal visits. For weeks 28-36, there should have been 4 visits. In the last 0.8 week, we can assume another prenatal visit. Thus, the total number of prenatal visits for high risk pregnancies is 7 + 4 + 1 or 12. These 12 visits translate roughly into a cost per visit of 1385 / 12 or about $115 for the control group.[295] A risk assessment visit would cost—we conservatively use a full 25% increase in cost—115 * 1.25 or approximately $143. These are figures from 1992-1994. Using 1993 as the average data collection point and converting via the Consumer Price Index, the result is approximately $175 in 2001. Compared to $180, such a cost allows the women to pay for one risk assessment visit and have a little money left over to pay for another such visit. It is highly unlikely that women paid for "partial" risk assessments. Thus, the average $180

[287] Ross, 835.
[288] Morrison, 43.
[289] Fangman, 749.
[290] Lear, 866.
[291] Hutti, 290.
[292] OB/GYN Nurse, Medical Department, MIT, telephone interview with author, March 31, 2005.
[293] Siran Koroukian and Alfred A. Rimm, "The 'Adequacy of Prenatal Care Utilization' (APNCU) Index to Study Low Birth Weight: Is the Index Biased?" *Journal of Clinical Epidemiology*, 55 (2002): 297.
[294] See Virtual Hospital, "Obstetrics: Prenatal Care," *University of Iowa Family Practice Handbook, Fourth Edition*, <http://www.vh.org/adult/provider/familymedicine/FPHandbook/Chapter14/02-14.html> (31 August 2005).
[295] We use control group costs, since by *concealing* their PHI these women cannot be enrolled in telemedicine.

WTP might suggest some women paid the $175 for one visit, while others paid for two or more of such visits to create the average WTP.

We must understand payment timing. *Which* risk assessments are paid for will dictate the cost implications to the insurer. If women paid for later risk assessments, PM would recognize their high-risk status earlier and enroll them in disease management; their WTP would have no effect on the insurer's operations. Based on our assumptions at the beginning of this discussion, that women would recognize the significance of such assessments, such women would want to pay for earlier risk assessment visits. They would want to protect privacy from the beginning. It would not make sense to disclose PHI and then subsequently try to protect PHI; privacy may not be protected. In the 2001 study, women would want to pay for the 12-week risk assessment, the first time they might learn more about their high-risk status. The women who pay more for privacy, as they create the average $180 WTP, may pay for the 24-week and any later risk assessment visits.

If risk assessments are performed in non-provider contexts, the women may act in other privacy "defensive" ways. For example, the California HealthCare Foundation Survey indicates that 2% of people nationwide decided not to be tested for a medical condition because they "were concerned that others might find out about the results."[296] Presumably the women would want to avoid seeing their providers or being subjected to any tests at the beginning of their condition to protect their privacy, not in the middle or in the end of their condition. As above, privacy would be protected from the outset. They may avoid the first or first few risk assessment visits.

2.2.4.3 Impact of Delayed Disease Management

The cost implications of such behavior to the insurer are that the women will be delayed in entering disease management. The intent of almost all high-risk population management programs is early detection.[297] [298] This was the purpose of disease management, as described before. In this case, since the first and in some cases the second risk assessment is paid for out of pocket or perhaps not conducted, such records should not be available to PM. With regard to out-of-pocket payments, data transmission between provider and insurer become limited because the provider is not seeking reimbursement. The women are paying the provider directly, lessening his need to submit any associated data to the insurer. In the worst case, telemedicine may not be administered at all. In Lear's 1998 study, two risk assessments were done by an HMO, before and roughly during the time when home uterine monitoring would typically be administered, at 24 weeks gestation. A high-risk pregnant woman would be enrolled in a

[296] California Healthcare Foundation, "Medical Privacy and Confidentiality Survey," 15.

[297] Jan Green, "Sizing Up the Sickest," *Hospitals & Health Networks*, 72 (1998): 28.

[298] Mary Anne Sloan, "Targeting Populations at Highest Risk: IT Delivers a 2:1 ROI for Midwest Health Plan – What Works: Disease Management," *Health Management Technology*, September 2003, <http://www.findarticles.com/p/articles/mi_m0DUD/is_9_24/ai_108148076> (20 April 2005).

case management intervention if either of the risk assessments indicated the woman was high risk. The women in Lear's study were designated as high- or low-risk based on the scores computed from the questions they and their physicians answered during the risk assessments. Lear shows how high-risk women with incomplete questionnaires, or low-risk women—those whose answered questions did not create a cumulative risk score designating them as "high risk"—could not be enrolled in case management. The HMO may not have understood the nature of their illness. Some of these women delivered preterm. In another study, preterm women who were not risk assessed by their provider delivered children who spent about four more days in the intensive care nursery as compared to children of preterm women who were risk assessed.[299] Based on the 2001 study, four days in the intensive care nursery can readily translate into thousands of dollars for the HMO.[300] Similar delays may happen in the context described by the 2001 study.

We look at the potential costs implicated due to the disease management delay. In the 2001 study, costs spanned from prenatal care to post-birth including any intensive care services used by the neonate. All the costs were categorized as prenatal care, antepartum hospitalization, delivery, intensive care nursery, and the disease management-like telemedicine services.[301] Costs starting at 24 weeks, the typical start of home uterine activity monitoring, and ending with the intensive care nursery would be implicated if the woman is not fully recognized to be high risk on time in our constructed telemedicine model. The 5 prenatal visits after the antepartum hospitalization, happening at close to 29 weeks gestation when home uterine activity monitoring was typically administered in the 2001 study, as described before, should be apportioned to the implicated costs as they were different for control and intervention groups. However, the costs of risk assessments done by the providers do not need to be so apportioned as costs should be approximately the same for control and intervention groups.[302] Keeping the same 7/12 fraction of the prenatal costs across control and intervention groups, the control group had an average delivery cost of about $20,000 while the intervention group had an average delivery cost of about $6400. The difference, the implicated costs, is about $13,000 in our constructed telemedicine model.

The impact to the insurer from women paying out of pocket or avoiding care appears to be a delay of the women's entry into disease management by one week. A week's worth of the implicated costs would be borne by the insurer. In the 2001 study, the total number of days for the infant in intensive care was about 8.7 days.[303] In total, the period wherein delays may lead to greater costs is approximately 98 days. This is 24 weeks for the initiation of home uterine activity monitoring in our designed telemedicine model subtracted from the 36.8 weeks plus 8.7 days for the length of the pregnancy and number of days spent by the infant in intensive care, respectively. During this time the child's condition may worsen, as shown in the 2001 study. Dividing the implicated costs,

[299] Fangman, 747.
[300] Morrison, 46.
[301] Morrison, 46.
[302] Implied OB/GYN Nurse, Medical Department, MIT.
[303] Morrison, 46.

$13,000, by 98 days, the incremental daily cost of delaying telemedicine is about $130. Note, this makes a somewhat unrealistic assumption of uniform distribution of costs across all the days. Actual daily costs can be lower or higher depending on the progress of the pregnancy. Several days or up to a week's delay into telemedicine seems appropriate. By that time the insurer should find if a woman is high risk.

There are other ways for the insurer to identify high-risk pregnant women. In the context described in the 2001 study, PM can use the third risk assessment record to identify the women. For those women who paid for only one or two visits via their WTP, data from the third risk assessment, at 30 weeks gestation, should be available because those women did not pay for that third visit. The insurer can also use methods not relying on risk assessments. Given the high expense generated by women giving birth preterm, insurance organizations should try to find such individuals early for intervention.[304] [305] [306] The women's physicians might refer the women to the insurer; the insurer can conduct a medical record review to find if pregnant women are high-risk; or the insurer can recognize a preterm labor hospitalization taking place for such women, as was done for the 2001 study.[307] [308] [309] [310] In the latter case, the monitoring staff, upon recognizing a hospitalization, can guess a pregnant woman may have preterm labor.[311] Not all these methods are optimal, as indicated before. For example, the insurer might be organizationally separate from the providers, limiting its ability to review medical records residing in provider offices. Nevertheless, if possible, all such methods could allow the insurer to enroll the woman into a telemedicine program. The preterm labor diagnosis and the third risk assessment which followed the preterm labor diagnosis chronologically in the 2001 study took place at approximately 5-6 weeks after the 24 weeks gestation when home uterine monitoring is typically initiated. We arbitrarily use a conservative 7 days for the delay into disease management, given the importance of identifying the women early, the failure of the insurer's processes to always detect high-risk women if risk assessments are not administered, and the steps the insurer might take to identify the women despite missing data. No study I found appears to assess the delay into telemedicine due to privacy concerns. I estimate a delay of 7 days given the described context, which will be examined in our sensitivity analysis; 130 * 7 or about $910 will be the cost to the insurer due to a single woman's WTP or avoidance of risk assessment visits.

[304] Morrison, 43.

[305] Melissa Muender, Mary Lou Moore, Guoqing John Chen, and Mary Ann Sevick, "Cost-benefit of a Nursing Telephone Intervention to Reduce Preterm and Low-birthweight Births in an African American Clinic Population," *Preventive Medicine*, 30 (2000): 271.

[306] Ross, 835.

[307] Fangman, 745.

[308] Ofman, 1607.

[309] Hutti, 291.

[310] Lear, 867.

[311] Typically women are hospitalized after symptoms of preterm labor. (See J. Sanin-Blair, M. Palacio, J. Delgado, F. Figueras, O. Coll, L. Cabero, V. Cararach, and E. Gratacos, "Impact of Ultrasound Cervical Length Assessment on Duration of Hospital Stay in the Clinical Management of Threatened Preterm Labor," *Ultrasound in Obstetrics & Gynecology*, 24 (2004): 756).

We compute the loss to the larger insurer using National Center for Health Statistics data. In 2001 there were 14.1 live births per 1000 population.[312] Of these, 11.9% were preterm births.[313] According to the 2001 study, on average about 40% of preterm births are due to preterm labor. Therefore, we have a total of 748,000 * 14.1 / 1000, or about 10,500 children being born. Of these, 10,500 * 0.119 or about 1240 will be born preterm; 40% of these or about 496 will be born preterm due to preterm labor. Thus, 496 * 0.0302 * 910 or about $13,600 will be the loss to the insurer due to lack of anonymity in the BQMA.[314] As this is 1992-1994 data, we again use 1993 as the average data collection point for computation. Converting via the Consumer Price Index, the 2001 result is about $16,600. The smaller insurer will have a proportionally smaller loss.

2.2.5 Reduction in Data Error

We demonstrate another case of "increase in operating costs" for the insurer due to lack of privacy protection. We quantify another application performance degradation to measure the intangible loss of weaker privacy protection. We show a financial loss to the insurer because there are errors in identifiers. Once again, the loss will not be large, but again it will be buttressed later on. The high-risk pregnancy analysis above will be used. In several studies on use of Social Security Numbers (SSN) within organizations, the Government Accountability Office found that some insurance organizations use the SSN as the primary identifier, which becomes the policy-holder's insurance number.[315] [316] A single identifier is apparently used for linking data. If an identifier is perceived error-free, using such an identifier becomes the easiest way to link records—it can be indexed and searched.[317] [318] Sometimes linkage identifiers are perceived to be credible and BQMA staff does not examine errors further, as suggested earlier. The BQMA linkage identifier will be referred to as the Medical Record Number (MRN). For the MRN error rate, we rely on the error rate given by a Health Plan Employer Data and Information Set

[312] Centers for Disease Control and Prevention. "National Vital Statistics Report: Births: Final Data for 2002," 4, <http://www.cdc.gov/nchs/data/nvsr/nvsr52/nvsr52_10.pdf> (26 April 2005).

[313] Centers for Disease Control and Prevention. "National Vital Statistics Report: Births: Final Data for 2002," 16.

[314] Note, in this analysis we assume each mother will only have one as opposed to multiple children. This is a reasonable assumption since in the US, national birth data show that only 3.3% of births are via "multiple gestation" (i.e., a mother giving birth to twins, triplets, etc.). (See Centers for Disease Control and Prevention. "National Vital Statistics Report: Births: Final Data for 2002," 98).

[315] Implied, General Accounting Office, "Social Security: Government and Commercial Use of the Social Security Number is Widespread," February 1999, 10, <http://www.gao.gov/archive/1999/he99028.pdf> (31 August 2005).

[316] Implied, General Accounting Office, "Social Security Numbers," January 2004, 12, <http://www.gao.gov/new.items/d04768t.pdf> (31 August 2005).

[317] For example, see William E. Winkler, "Preprocessing of Lists and String Comparison," in W. Alvey and B. Kilss (eds.), *Record Linkage Techniques*, 1985, U.S. Internal Revenue Service, 181.

[318] Department of Public Health, State of Massachusetts staff, email to author (LinkPro 2.0 documentation), July 18, 2002.

(HEDIS) auditor.[319] The National Committee for Quality Assurance, which administers HEDIS, recommends such individuals for various HEDIS oversight functions. In an interview with one HEDIS auditor, he mentioned that, of the many audits he's performed, an internal MRN error rate of 1% seemed reasonable. Given our discussion of a possibly higher error rate in Section 1.3.2, as well as the similarity of the BQMA, we use such an error rate for our analysis.

A 1% MRN error rate will have an impact similar to but more complex than that of the high-risk pregnancy case above. If a PM record representing the first or second risk assessment is in error, there will be a delay in administering telemedicine. PM will not find that record; however, it should find the record for the third risk assessment visit, which should be available if that visit wasn't paid for out of pocket, as described above. Alternatively, the insurer can rely on the other methods mentioned above, such as recognizing the preterm labor hospitalization taking place, to find the high-risk pregnant women. Given a 1% MRN error rate, the loss from a single risk assessment record in error is 910 * 0.01 or about $9.10. This loss must be doubled, as both risk assessment records arrive to the insurer at different earlier times, thus subject to different independent earlier mistakes. For the larger insurer, the loss becomes 2 * 496 * 9.1 or roughly $9000 using 1992-1994 data. Converting to 2001 for consistency, the loss becomes about $11,000. If the third risk assessment record itself is in error, or if other records acquired via non-risk assessment methods are in error, too, costs will be higher. Assuming these records are maintained electronically, PM will have to wait for additional electronic data to enroll the disease management candidates. The smaller insurer will have a proportionally smaller loss.

2.2.6 Cost of Deidentification Technology

We now quantify the costs associated with providing BQMA privacy protections. In the technical part of this thesis, we will attempt to deidentify the data used by the BQMA while permitting the applications to function, as the method for privacy protection. We will rely on HIPAA's deidentification standard known as Safe Harbor. We will provide the costs associated with applying Safe Harbor to the copy data store in this section and examine the actual technology to deidentify the data in the next chapter.

We use costs from several available studies to quantify the expenses of data deidentification. The First Consulting Group created a cost model in 2000 for the American Hospital Association when the Group was estimating the cost of future HIPAA compliance for US hospitals.[320] One cost estimated by the First Consulting Group was the implementation of HIPAA's "minimum necessary use" requirement, which requires

[319] Charles Chapin, a HEDIS auditor connected to the National Committee for Quality Assurance, telephone interview with author, February 22, 2005.

[320] HospitalConnect, "The Impact of the Proposed HIPAA Privacy Rule on the Hospital Industry," December 2000, <http://www.hospitalconnect.com/aha/key_issues/hipaa/content/FCGDecember2000.pdf> (31 August 2005).

removing access to data fields from individuals who do not need to know the data for their daily work. An actuary who is computing premiums, for example, does not need to know physician identifiers or patient names to compute the premiums as they should not be used in the computations. We can use the values provided by the First Consulting Group for the "minimum necessary use" change to quantify BQMA deidentification costs. The nature of obfuscation under "minimum necessary use" should be similar to that of Safe Harbor as sensitive variables are removed or modified to hide information about the represented people, as will be shown in chapter three. For the sake of this thesis, we'll take the representative costs for implementing HIPAA's "minimum necessary requirement" for applications in a hospital IT system as the costs to change the BQMA used by an insurance organization. In fact, we think this might be an overestimate since hospital systems might be more complex. Nevertheless, some of the hospital IT systems had similar functionality to that of the BQMA. The First Consulting Group identified that, on average, each hospital had 17 major different IT subsystems which required upgrades.[321] Two of the 17 subsystems were the hospital Utilization Review and Case Management software platforms. A hospital's Utilization Review system, for example, might be similar to an insurer's Utilization Review system, requiring similar changes to the applications.[322] The hospital's Case Management application may again be similar to the insurer's PM and Disease Management programs, requiring similar modifications to the associated software applications.[323] [324]

We compute the costs to deidentify one BQMA. The First Consulting Group estimated an initial cost of $15.79 per employee for staff training and $0.94 per employee for on-going annual training related to the "minimum necessary use" requirement for each hospital across the 6050 US hospitals the Group examined. The Group also estimated an initial cost of $17,395 per hospital to plan for the compliance with the "minimum necessary use" requirement across the 6050 US hospitals. The Group estimated an annual $9073 per hospital to monitor for compliance with the "minimum necessary use" change as well. The IT changes required for the "minimum necessary use" requirement across each of the 6050 US hospitals were estimated to range between $142,452 and $3,175,232. The Group also estimated annual operating costs of between 0 and $7167 per hospital for maintaining the software necessary for the "minimum necessary use" tenet. Thus, on average, the initial IT modification costs for the "minimum necessary use" change become about $1,658,840 per hospital, assuming a normal distribution of costs across the 6050 US hospitals. The annual IT operating expenses become about $3583 per hospital, again assuming IT maintenance costs are normally distributed across the 6050 hospitals. Converting to 2001 figures, the First Consulting Group costs above become about $16.23 per employee for the initial training; about $0.97 per employee for the annual re-training; roughly $17,879 for initial compliance planning; about $9325 for annual compliance

[321] HospitalConnect, "Report on the Impacts of the HIPAA Final Privacy Rule on Hospitals," March 2001, <http://www.hospitalconnect.com/aha/key_issues/hipaa/content/FCGMarch2001.doc> (31 August 2005).

[322] Center to Advance Palliative Care (CAPC), "Utilization Review," *CPAC Manual*, 20 February 2002, <http://64.85.16.230/educate/content/development/utilizationreview.html> (31 August 2005).

[323] Case Management Society of America (CMSA), "CMSA Definition and Philosophy," <http://www.cmsa.org/AboutUs/CMDefinition.aspx> (31 August 2005).

[324] Case Management Society of America, "Strategic Vision," <http://www.cmsa.org/PDF/StrategicVision.pdf> (31 August 2005).

monitoring; about $1,705,070 for the initial IT modifications; and about $3682 for the annual IT maintenance effort for the "minimum necessary use" change. From Section 2.2.2, we estimate it takes 4 individuals to operate one BQMA. Therefore, the initial training cost becomes 16.23 * 4 or about $64.92 for the "minimum necessary use" change. The annual re-training cost becomes 0.97 * 4 or about $3.88 for the modification. We divide the other four costs by 17 to obtain a per-system cost. Thus, an insurer would pay about $1050 for the initial planning; about $548 for the annual compliance monitoring; about $100,000 initially for the IT modifications; and roughly $216 to annually maintain the IT modifications. All these costs would be the approximate costs to deidentify one BQMA.

Deidentifying four BQMA would imply quadrupling these amounts. However, total costs may be less if all four BQMA rely on the same data, e.g., one copy data store is used to run all four BQMA. If the same staff performs the deidentification, costs should be reduced because the staff would not have to re-learn the deidentification approaches as it tackles another BQMA. Faster learning by the relevant staff may quicken the needed IT changes. We can arbitrarily double the IT-related costs of deidentifying a single BQMA to quantify the IT-related costs of deidentifying the four applications. We have 100,000 * 2 or about $200,000 initially and about 216 * 2 or $432 annually as the initial IT modification and subsequent maintenance costs for creating and operating the four deidentified BQMA. We multiply the costs related to staff training and re-training and the costs of compliance planning and compliance monitoring by four since these activities may have to be done independently for each BQMA. The purpose of each of the BQMA is relatively different, as described earlier, hence, probably requiring different training and compliance activities regarding each of these applications. Thus, we have a cost of (1050 + 64.92) * 4, or approximately $4450, for initial BQMA staff training and compliance planning purposes. We also have a cost of (548 + 3.88) * 4, or about $2200, for annual staff re-training and compliance monitoring purposes. The combined costs for deidentifying the four BQMA become 200,000 + 4450, or about $204,000 initially, and 432 + 2200 or about $2630 annually. This is about 2 full-time equivalent (FTE) employees working for one year plus a small percent of the work of one FTE employee working annually thereafter.[325]

2.2.7 Net Benefits to the Health Insurer

We compute the net benefits to the insurer from installing our privacy-protecting approach. We ignore the benefit when policy-holders switch health plans. Privacy protection may not create a long-term gain as the smaller insurer—or larger insurer if the smaller insurer installs privacy protections first—can also install safeguards to protect privacy. Policy-holders might return to the smaller (or larger) insurer. We include the $320 gain for the insurer as the insurer can annually benefit from preventing staff from

[325] Based on Meghan Dierks, M.D, Instructor in Medicine, Harvard Medical School, personal discussion with author on November 22, 2005; Hal Abelson, Professor, Electrical Engineering and Computer Science, MIT, personal discussion with author on November 22, 2005.

abusing identifiable PHI. We ignore the roughly $4 million for which the larger insurer may be liable due to the willingness-to-pay (WTP) for privacy protection, since that money should be returned to the insurer as the insurer might recapture the WTP through various recovery mechanisms as described before. We include both benefits related to high-risk pregnancies—the roughly $16,600 related to improved disease management outcomes and the roughly $11,000 related to minimizing MRN errors—as the insurer can annually benefit from improved privacy protection within both operations. The annual benefit to the larger insurer becomes 320 + 16,600 + 11,000 or about $27,900. After nine years, the benefits should cover the costs of implementing our deidentification approach. We use a discount rate of 3%.[326] The net present value of a constant annual revenue stream of $27,900 over 9 years is (27,900/0.03) * [1 – (1/(1 + 0.03)^9)], or about $217,000.[327] The net present value of the combined installation and operation costs of the deidentification approach after 9 years is (204,000/(1 + 0.03)) + (2630/0.03) * [1 – (1/(1 + 0.03)^8)] or about $216,000. We will explore later on how payback should happen considerably sooner given that the insurer suffers other key opportunity costs as will be discussed in subsequent sections. The smaller insurer should have a longer payback period because its benefits are smaller. Of course, the insurance organization can absorb the cost of deidentification much as in the $4 million WTP discussion from before. We assume the insurer wants to recover costs for our analysis as a conservative worst-case assumption.

2.2.8 Sensitivity Analysis

2.2.8.1 Health Plan Switching

Referring back to our original economic decision model, we now perform step 7, the sensitivity analysis. The sensitivity analysis on the core components of the net benefits computation above—combining the gains from reducing losses due to identifiable PHI access, improving disease management, and reducing identifier errors, with the costs of the deidentification—will be performed. In addition, a sensitivity analysis on the non-core components of the net benefits computation—the gain from health plan switching behavior and willingness-to-pay-related dynamics—will also be conducted. If some non-core net benefit items are at least partially present in some US health insurance marketplaces, the sensitivity analysis will point to the key parameters influencing those values.

The key financial drivers affecting the core and non-core components are examined, presented in the same order as these components were described in Sections 2.2.1 through 2.2.6.[328] [329] [330] First, we examine competitive profitability. A key parameter

[326] Hunink, 276.

[327] See Hunink, 273.

[328] Hunink, 344.

affecting the financial benefit to the insurer due to health plan switching is the size of the smaller insurance organization. In this analysis, assume the larger insurer installs privacy protections first. A similar analysis can be carried out if the smaller insurer incorporates such protections first. If the smaller insurer is relatively large then the number of new policy-holders the larger insurer gains will also be relatively large because the number of "switchers" is a percentage of total members. Assuming that acquiring more policy-holders leads to profitability, as discussed before, the larger insurer should become more profitable when it installs better privacy-protective practices. Consider the state of California. Its "smaller" insurance organizations are among the largest in the US. Based on the cost model we're creating, it's possible that California's largest health insurer could see a very large profit from installing privacy-protective practices. California had the largest number of individuals insured by the workplace, approximately 17,791,795 people in 2001.[331] 0.6 * 17,791,795 or about 10,600,000 Californians had a choice of at least two health plans.[332] The top group health insurer had a 31% market share while the top 3 group health insurers had a combined 61% market share.[333] Assume the second top insurer had a (61-31)/2 or 15% market share for analysis, for simplicity, assuming the second and third largest insurers had equal market shares in California. During annual re-enrollment, the larger insurer would enroll 10,600,000 * 0.31 or roughly 3,280,000 policy-holders. The smaller insurer would enroll 10,600,000 * 0.15 or about 1,590,000 policy-holders. If the larger insurer installed privacy-protecting practices, 0.036 * 1,590,000 or about 57,200 policy-holders would switch to the larger insurer. The larger insurer would gain 57,200 * 61.8 or about $3,530,000 annually while the smaller insurer would lose this amount in 2003. Converting to 2001, the profit and loss would become $3,300,000, respectively.

Compare this value to that of insurers in North Dakota. North Dakota had much smaller "smaller" health insurance organizations in 2001. If the largest North Dakota insurer installed privacy safeguards first, it would profit considerably less than California's insurers. In 2001, approximately 344,379 individuals were insured through their workplace in North Dakota.[334] Of this figure, 60%, or about 206,000 individuals, could choose from at least two health plans. The top insurer had a 91% market penetration compared to the 96% penetration of the largest 3 health insurers.[335] Assume, based on

[329] Samuel Wang, Blackford Middleton, Lisa A. Prosser, Christiana G. Bardon, Cynthia D. Spurr, Patricia J. Carchidi, Anne F. Kittlera, Robert C. Goldszer, David G. Fairchild, Andrew J. Sussman, Gilad J. Kuperman, and David W. Bates, "A Cost-benefit Analysis of Electronic Medical Records in Primary Care," *The American Journal of Medicine*, 114 (2003): 400-401.

[330] Luisa Franzini, Elena Marks, Polly F. Cromwell, Jan Risser, Laurie McGill, Christine Markham, Beatrice Selwyn, and Carrie Shapiro, "Projected Economic Costs Due to Health Consequences of Teenagers' Loss of Confidentiality in Obtaining Reproductive Health Care Services in Texas," *Archives of Pediatrics & Adolescent Medicine,* 158 (2004): 1143.

[331] The Henry J. Kaiser Family Foundation, "Health Insurance Coverage in America, 2001 Data Update," 33.

[332] Note, in this discussion we again assume children are some of the insurer's members which will participate in the plan switching, via their parents switching, as referenced before.

[333] AcademyHealth, "Mapping State Health Insurance Markets, 2001: Structure and Change," 13.

[334] The Henry J. Kaiser Family Foundation, "Health Insurance Coverage in America, 2001 Data Update," 33.

[335] AcademyHealth, "Mapping State Health Insurance Markets, 2001: Structure and Change," 13.

this 5% difference, the insurer with the second largest penetration had a (96-91)/2 or 2.5% market share for analysis. We assume, as above, the second and third largest insurers had equal market shares in North Dakota, for simplicity. The larger insurer would enroll 0.91 * 206,000 or about 187,000 policy-holders while the smaller insurer would enroll 0.025 * 206,000 or about 5150 members in 2001. If the larger insurer would incorporate privacy-protective practices first, it would gain an annual 0.036 * 5150 * 61.8 or about $11,400 profit in 2003. The profit would be about $10,900 in 2001. The smaller insurer would lose such an amount annually.

2.2.8.2 Removing Identifiable Data Access

We perform a sensitivity analysis when analyzing the removal of identifiable PHI access. The most important factor affecting insurer gain is the total number of employees per organization. Gains are smaller with increasing number of employees as per-employee losses are diluted. This is particularly shown in the CSI/FBI survey. When examining the East Coast malpractice data set, several organizations with the smallest number of employees per organization each had a loss of zero across each of the years of the data set. This implies no gain to the insurer from adding privacy protections to prevent employees from misusing identifiable PHI because the insurer is not losing money from identifiable PHI access.

The 2001 CSI/FBI survey reports higher losses across its respondents. Using the smallest number of employees per organizational-size category, the CSI/FBI survey suggests 16% of respondents had 1 employee; 16% had 100; 8% had 500; 22% had 1000; 11% had 5001; and 27% had 10,000.[336] The average number of employees per respondent becomes about 3500. The gain to the insurer from removing identifiable PHI access becomes (2,210,000 / 3500) * 16 or about $10,000. Hence, the total gain to the insurer from deidentifying the BQMA PHI when number of employees is small is about 10,000 + 0, or about $10,000.

Compare such values to the case when the number of employees is large. The East Coast malpractice data set shows that for organizations with some of the largest number of employees, across all the years of the data set the loss of each such organization was less than a dollar per employee or less per year. The CSI/FBI survey suggests similar low values. We use the highest number of employees per organization from the 2001 Fortune 500 list, 1,383,000, for a conservative analysis.[337] Using the maximum number of employees in the number-of-employees category, 16% of the CSI/FBI survey respondents had 99 employees; 16% had 499; 8% had 999; 22% had 5000; 11% had 9999; and 27% had 1,383,000. These compute to an average number of about 370,000 employees per organization. The total loss to the larger insurer for the four BQMA from the CSI/FBI data is (2,210,000 / 370,000) * 16 or at most $96. Thus, the benefit to the

[336] Computer Security Institute, "2001 CSI/FBI Computer Crime and Security Survey," 3.

[337] "Fortune 500 Largest US Corporations," F-31.

insurer from deidentifying the BQMA PHI when number of employees is large is at most 16*1 + 96 or approximately $112.

2.2.8.3 Paying-out-of-pocket Dynamics

We conduct a sensitivity analysis on the paying out-of-pocket dynamics. The most important parameter is the type of insurance organization involved. Traditional fee-for-service plans require less PHI, while full managed care organizations require more PHI for administration and oversight. The latter organizations generate a higher willingness-to-pay as individuals are more concerned about potentially intense BQMA PHI use. We vary the percent of individuals who don't submit claims based on their insurance type. In traditional plans, the willingness-to-pay is close to zero due to fewer concerns over BQMA PHI as the BQMA are not especially prevalent within such organizations. In the most restrictive managed care organizations, 10.8% of policy-holders would pay out of pocket 748,000*180*(0.108) or about $14,500,000 rather than submit claims to the larger insurer due to concerns about persistent BQMA use. The smaller insurer will face a significantly smaller total maximum willingness-to-pay as it has fewer policy-holders.

2.2.8.4 High-risk Pregnancy Assessment

A sensitivity analysis of the disease management program for high-risk pregnancies shows that the main driver of the insurer's gain is the amount of delay before a woman enrolls in telemedicine. The longer the delay, the greater the benefit to the insurer from adding technologies and policies to protect privacy as the delay can be reversed. If the delay is close to zero, the financial benefits of installing privacy-protecting practices to the insurer could be close to zero. If the insurer uses some of the methods we described in earlier discussions, such as if its staff waits for a hospitalization related to preterm labor and enrolls the women in telemedicine, without their "cooperation," there may be no delays. It may not make sense for the insurer to install privacy-protecting practices since the delays can hardly be reduced further. On the other hand, telemedicine may not be administered at all and the number of days delayed can be maximal. The financial benefit to the insurer from installing additional privacy protections becomes 496 * 0.0302 * 130 * 98 or about $190,000 in 1992-1994 as installing such protections would enroll the women without the delays. Converting to 2001, using 1993 as the average year of costs, the benefit becomes about $232,000. The smaller insurer would have considerably smaller benefits.

2.2.8.5 Medical Record Number Errors

A sensitivity analysis on errors in linkage identifiers suggests that the primary factor responsible for the insurer's gain is whether more than one variable is used for linkage. If more than one is used, the gain to the insurer can drop dramatically. Other linkage variables can be used to identify the same individuals if some linkage variables are in error. The gain to the insurer from installing privacy-enhancing technologies may be close to zero because errors hardly exist in linking policy-holder records.

2.2.8.6 Variability in Deidentification Cost

A sensitivity analysis on the cost of BQMA deidentification shows that the complexity of the BQMA is the main driver of the insurer's deidentification costs. The larger the programming effort for the IT-related changes, clearly, the more expense. The BQMA may be sufficiently sophisticated so that the IT-related deidentification changes do not involve extensive change. For example, the changes I propose to deidentify the BQMA include encrypting fields instead of revoking access to them, as we will see later on. Many recent versions of commercial databases allow for column level encryption of fields, which should reduce implementation expense.[338] [339] [340] [341] Performing the encryption may be done without the potential cost of installing additional encryption software. BQMA systems may also be less complex so that programming to create the deidentification changes would be more straightforward than for more involved system architectures. We can use the lowest IT-related costs suggested by the First Consulting Group to find the lowest cost for BQMA deidentification. From Section 2.2.6, the lowest IT-related cost to initially create the "minimum necessary use" modifications is $142,452 per hospital in 2000. No additional money is needed for annual maintenance costs for this change. Converting to 2001, this cost becomes approximately $146,422. The IT-related deidentification cost of the four BQMA becomes an initial sum of 2 * 146,422 / 17 or roughly $17,200. This must be combined with the training and compliance monitoring costs from before. Therefore, we have an initial cost of 17,200 + 4450 or about $21,600 to create the BQMA deidentification changes. There is also a cost of 0 + 2200 or $2200 to manage the deidentification changes annually.

[338] Ecommerce Times, "Top Dog Oracle Losing Database Market Share," 11 March 2003, <http://www.ecommercetimes.com/story/20968.html> (31 August 2005).

[339] Database Journal, "SQL Server 2005 Security - Part 3 Encryption," 22 February 2005, <http://www.databasejournal.com/features/mssql/article.php/3483931> (31 August 2005).

[340] Jared Still, "Data Obfuscation and Encryption," <http://www.cybcon.com/~jkstill/util/encryption/data_obfuscation_and_encryption.html> (31 August 2005).

[341] IBM, "Cost of Encryption for DB2," *IBM DB2 Tools*, <http://publib.boulder.ibm.com/infocenter/dzichelp/index.jsp?topic=/com.ibm.imstools.deu.doc.ug/cost.htm> (31 August 2005).

Conversely, costs might be larger if system changes are difficult or systems need to be replaced to obtain the needed functionality. We use the highest First Consulting Group estimate for IT-related costs, from Section 2.2.6, of $3,175,232 to install and $7167 to maintain, per hospital, the 17 major different hospital IT systems in 2000. Converting to 2001, these costs become approximately $3,263,740 and $7366, respectively. Dividing by 17 and multiplying by 2 we obtain about $380,000 and about $866 as the IT-related costs to create and operate the changes for the four BQMA, respectively. Again, adding the non-IT costs to these expenses, we have an initial creation cost of 380,000 + 4450, or about $384,000, and an annual cost of 866 + 2200, or about $3060, to manage the BQMA deidentification changes. The smaller insurer may have costs slightly lower than the highest possible larger insurer costs as systems of the smaller insurer may be less complex.

2.2.8.7 Sensitivity Analysis Conclusion

Under a few sensitivity scenarios, the insurer can cover the costs of BQMA deidentification within a short amount of time, possibly within a few years. Benefits from preventing BQMA staff from accessing the PHI range from $112 to $10,000. Benefits from enrolling more high-risk pregnant women range from $0 to $232,000. Benefits from addressing the errors in the medical record numbers to improve telemedicine administration range from $0 to $11,000. The costs to deidentify the BQMA range from $21,600 to $384,000 for the initial tasks to create the changes. The annual cost to manage the changes may be at most $3060. Imagine that the larger insurer experiences the lowest benefits. The insurer has other reliable methods to quickly identify high-risk pregnant women without using risk assessments. It uses multiple reliable variables for BQMA linkage. Women can be readily identified despite non-cooperation and data within the copy data store can be accurately linked. The insurer would acquire very low financial benefits from privacy protections as the context is hardly improved from such protections. Now imagine that the larger insurer experiences the highest benefits. The insurer has few other reliable methods to identify high-risk women aside from the risk assessments. The insurer is primarily using only one linkage variable within the BQMA. High-risk pregnant women cannot be readily located and data linkage is not optimal. The overall annual benefit to the insurer becomes 10,000 + 232,000 + 11,000 or about $253,000. The sum of the benefits that would accrue to the insurer would cover the costs of BQMA deidentification in less than two years. Twice the annual insurer benefit is 2 * 253,000 or $506,000. The maximal BQMA installation and operating cost for the first two years would be about 384,000 + 3060 or about $387,000—less than the $506,000. Achieving profitability would take longer for the smaller insurer as it acquires less benefits.

2.2.9 Additional Financial Benefits

The financial benefits of implementing additional BQMA privacy protections are strengthened by still other gains, suggesting time for return on investment can be further decreased. First, there are cost implications to the insurer from additional disease management programs. We make the assumption that the insurer would adopt a PM platform for all its disease management protocols to effectively identify chronically ill or high-risk individuals. Missed cost savings could happen through an HIV disease management program. One insurance-based disease management program showed the cost savings of treating HIV positive patients.[342] [343] The insurer would want to adopt a similar program to lessen the costs of managing HIV positive policy-holders. HIV is obviously a privacy-sensitive condition.[344] Gains would be lessened if HIV patients pay out of pocket, avoid care, or there are errors in linkage identifiers because the associated PM software may not find the disease management candidates' data.

We quantify the percent of an insurer's policy-holders who have the confidentiality-sensitive condition of high-risk pregnancy or HIV who would reduce the disease management savings to the insurer, to understand the percent of individuals involved. We will apply national prevalence statistics as well as statistics from the larger insurer computed earlier which were based on national prevalences, to obtain the prevalence within a typical insurer, of women who will deliver preterm due to preterm labor and of people who are HIV-positive. In summary, across its entire population, the large insurer should have 496 / 748,000 or about 0.066% of policy-holders who are females who would deliver preterm due to preterm labor. In 2000, there were also approximately 900,000 individuals living with HIV in the US.[345] This represents roughly 0.31% of the 2000 US population.[346] Thus any insurer today, in 2005 or 2006, would have to manage roughly 0.31% of all of its policy-holders who have HIV, assuming the same US HIV prevalences. I acknowledge that there may be a disproportionate number of HIV-positive patients who are not privately insured and that a for-profit insurer may not have so many of such patients. Nevertheless very roughly, 3.02% of the above policy-holders, or very roughly 0.0302 * (0.0031 + 0.00066) or 0.011% of policy-holders, may pay out of pocket or avoid care creating poorer financial outcomes for an insurer because the insurer cannot optimally administer disease management. The assumption is that there is no overlap

[342] State of Florida, Agency for Health Care Administration, "The Florida Medicaid Disease Management Experience," 26 January 2005, <http://www.fdhc.state.fl.us/Medicaid/deputy_secretary/recent_presentations/medicaid_disease_management_house_012605.pdf> (31 August 2005).

[343] Robert Catalla, F.S. Goldstein, and C. Farthing, "The Disease Management Initiative – A Novel Approach in the Care of Patient with HIV/AIDS" (a poster presentation at the United States conference on AIDS, September 2001).

[344] American Civil Liberties Union, "ACLU Says CDC Guidelines on HIV Surveillance Could Lead to Better Privacy Protections," *News*, 1999, <http://www.aclu.org/Privacy/Privacy.cfm?ID=8791&c=27> (Jul 17, 2005).

[345] "Diagnoses of HIV/AIDS—32 States, 2000-2003," *Morbidity and Mortality Weekly Report*, 53 (2004): 1106-1110.

[346] US Census Bureau, "National and State Population Estimates."

between women who will deliver preterm due to preterm labor and people who are HIV positive, otherwise the computed percent above would be smaller.

Identification errors would lead to missed financial opportunities regarding the insurer enrolling patients in the disease management programs for asthma, diabetes, lower back pain, as well as those handling multiple chronic conditions simultaneously including the just-mentioned conditions plus coronary artery disease and congestive heart failure.[347] The insurer may want to adopt such programs. Identifier errors would prevent PM from linking data and identifying candidates in a timely manner. We again use national prevalence statistics and assume that the prevalence of such chronic conditions among an insurer's policy-holders is the same as national prevalence values. Using 2003 prevalence data, 7.0% of a typical insurer's policy-holders would have asthma.[348] Per 2005 data, any insurer should also have approximately 7.0% of individuals with diabetes.[349] In 2002, an estimated 4.8 million Americans had congestive heart failure.[350] This represented about 1.6% of the 2002 population.[351] In 2003, the prevalence of coronary artery disease (CAD) was approximately 13.2 million in the US population.[352] Based on the 2003 US population, the prevalence of CAD was approximately 4.5%.[353] Few consistent statistics appear available for the prevalence of chronic back pain in the US population. We used a 1995 North Carolina study, indicating a chronic low back pain prevalence of 3.9% in the North Carolina population, as an estimate of prevalence for the US population.[354] Therefore, 1% of policy-holders with these conditions, or 0.01 * (0.07 + 0.07 + 0.016 + 0.045 + 0.039) or about 0.24% of an insurer's members with these conditions will be affected due to errors in identifiers. Again, the assumption is that there is no overlap between the people affected by the different conditions, otherwise the computed percent would be smaller. Disease management candidate identification or risk stratification problems regarding such conditions may reduce cost savings to the insurer.

Additional savings to the insurer arise because the insurer faces opportunity costs regarding its employer clients who are also losing money due to suboptimal disease

[347] American Association of Health Plans/Health Insurance Association of America, "The Cost Savings of Disease Management Programs: Report on a Study of Health Plans."

[348] Centers for Disease Control and Prevention, "Current Asthma Prevalence Percents by Age, United States: National Health Interview Survey, 2003," 9 March 2005, <http://www.cdc.gov/asthma/NHIS/2003_Table_4-1.pdf> (10 March 2006).

[349] Centers for Disease Control and Prevention, "National Diabetes Fact Sheet," *Publications and Products*, 16 November 2005, <http://www.cdc.gov/diabetes/pubs/estimates05.htm#prev> (10 March 2006).

[350] Centers for Disease Control and Prevention, "Congestive Heart Failure and Adrenergic Receptor Polymorphisms," *Genomics and Disease Prevention*, 27 November 2002, <http://www.cdc.gov/genomics/hugenet/ejournal/heartfailure.htm#2> (10 March 2006).

[351] US Census Bureau, "National and State Population Estimates."

[352] See Thomas Thom, Nancy Haase, Wayne Rosamond, Virginia Howard, "Heart Disease and Stroke Statistics – 2006 Update," *Circulation*, 113 (2006): e86, e100. Note, the prevalence stated in this article actually refers to prevalence of coronary heart disease (CHD). However, typically, in public health discussions, references to CAD usually refer to CHD. (Nancy Haase, Biostatistics Program Coordinator, American Heart Association, telephone interview with author, March 28, 2006). Therefore, we can use CHD prevalence in this article to obtain the CAD prevalence needed in the text.

[353] US Census Bureau, "National and State Population Estimates."

[354] TS Carey, Evans A, Hadler N, Kalsbeek W, McLaughlin C, Fryer J, "Care-seeking among individuals with chronic low back pain," *Spine*, 20 (1995): 312-7.

management administration. Implementing additional BQMA privacy protections could convert the opportunities costs into financial gains for the insurer. Chronic and high-risk conditions increase the costs to employers, such as via employees' productivity declines, as employers must manage their employees' worse health.[355] [356] [357] In the interest of reducing such productivity and financial impacts to their organizations, employers may pay the insurer to improve the care of their "high-risk" employees. The insurer might solicit such payments. Using their leverage, some large employers and some employer purchasing consortiums have recently created financial incentives to encourage health plans to provide quality care to their employees.[358] Such incentives have included providing financial bonuses to health plans that have met or exceeded target employee health care metrics. The insurer can improve employees' care by enrolling more individuals in disease management by implementing additional privacy protections. Based on current precedent in the marketplace, the insurer might ask for the employer to share some of its productivity returns with the insurer for the latter's efforts.

Furthermore, the insurer faces opportunity costs due to losses from poorer data quality in the other BQMA and even non-BQMA software, which can also be converted to financial gains for the insurer. Losses to the insurer can arise from poor data within the other three BQMA. Poor data quality within the Utilization Review, Provider Profiling, and the Health Plan Employer Data and Information Set applications may also undermine the cost savings that such applications provide to the insurer. Assuming some of these provide cost savings for the insurer, creating additional BQMA privacy protection would provide the insurer financial benefits. By improving the other BQMA's data quality those applications' operations improve. Non-BQMA applications would also operate better because claims data would be available for processing throughout the organization. Formerly missing claims data would now be *submitted* to the insurer by policy-holders. This would improve the operations and thus cost-efficiency of the insurer's non-BQMA applications relying upon that data.

2.3 Organizational Support

Providing additional BQMA privacy safeguards may also lead to better identification of policy-holders for disease management which may then lead to the policy-holders' better care, another motivation within our technology adoption model. Regarding the analysis of premature pregnancies, providing extra BQMA privacy protection may reverse the women's out-of-pocket payment behavior. This may eliminate their 7-day delay into

[355] G. G. Liu, D. Ying, and R. Lyu, "Economic Costs of HIV Infection: An Employer's Perspective," *The European Journal of Health Economics*, 3 (2002): 226.
[356] Wayne Burton and Catherine M. Connerty, "Worksite-based Diabetes Disease Management Program," *Disease Management*, 5 (2002): 1-2.
[357] Chris Penttila, "An Ounce of Prevention…," *Entrepreneur Magazine*, January 2003, <http://www.findarticles.com/p/articles/mi_m0DTI/is_1_31/ai_n13470627> (31 August 2005).
[358] AcademyHealth, "Ensuring Quality Health Plans: A Purchaser's Toolkit for Using Incentives," 14, <http://www.academyhealth.org/nhcpi/healthplanstoolkit.pdf> (31 August 2005).

disease management due to better PM identification, which in turn can reverse the suboptimal care provided to the women and their newborns. When analyzing the characteristics of preterm children, researchers typically focus on a neonate's low birth weight, often at 2500 grams and below; a baby's admission to the neonatal intensive care unit (NICU); or if a baby was delivered before 37 weeks gestation.[359] [360] [361] Such characteristics of the neonate are associated with the neonate's higher morbidity and mortality, such as getting chronic lung disease, severe brain injury, retinopathy of prematurity, and neonatal sepsis.[362] In later years, preterm infants are at increased risk of motor and sensory impairment and behavioral problems.[363] [364] In the 2001 study, women not enrolled in telemedicine differed from those enrolled in telemedicine across these three dimensions: 1) neonatal birth weight for the control group averaged 2554 grams, and almost 60% of these neonates weighed below 2500 grams, while for the intervention group birth weight averaged 3224 grams; 2) neonates whose mothers did not receive telemedicine spent an average of 7.2 days at the NICU, while those whose mothers received telemedicine spent an average of 0.3 days at the NICU; 3) the average gestational age at delivery for the control group was 35.3 weeks, while for the intervention group the average gestational age at delivery was 38.2 weeks.[365] A delay of 7 days in getting telemedicine implies that 496 * 0.0302 or about 14.9 children would begin to experience the poor clinical and, later, sensory and behavioral outcomes within the larger insurer. The errors in identifiers suggest that an additional 2 * 496 * 0.01 or about 9.9 children would also experience these outcomes within the larger insurer.

The quality of care of policy-holders with other chronic conditions may decline due to confidentiality and identifier errors. Incorporating additional BQMA privacy protections would reverse the suboptimal administration of disease management regarding such conditions. People's defensive behavior could reduce the quality of their care if they have HIV and are not properly enrolled in disease management.[366] Similar outcomes would result with regard to depression, a confidential condition, which also has disease management protocols.[367] Errors in identifiers will also impact the care of individuals with these conditions as well as of those with less confidentiality-sensitive conditions which also have disease management protocols, including diabetes, asthma, congestive heart failure, chronic obstructive pulmonary disease, coronary artery disease, being frail

[359] Morrison, 46.

[360] Ross, 839.

[361] Corwin, 1284.

[362] Stavros Petrou, Ziyah Mehta, Christine Hockley, Paula Cook-Mozaffari, Jane Henderson, and Michael Goldacre, "The Impact of Preterm Birth on Hospital Inpatient Admissions and Costs during the First 5 Years of Life," *Pediatrics*, 112 (2003): 1290.

[363] Stavros Petrou, 1290.

[364] Shoo Lee, Douglas D. McMillan, Arne Ohlsson, Margaret Pendray, Anne Synnes, Robin Whyte, Li-Yin Chien, and Joanna Sale, "Variations in Practice and Outcomes in the Canadian NICU Network: 1996-1997," *Pediatrics*, 106 (2000), 1070-1079, <http://pediatrics.aappublications.org/cgi/content/full/106/5/1070> (1 September 2005).

[365] Morrison, 46.

[366] See similar concepts in H.B. Krentz, "The High Cost of Medical Care for Patients Who Present Late (CD4 < 200 cells/uL) with HIV Infection," *HIV Medicine*, 5 (2004): 93-98.

[367] Welch, 356.

and elderly, cancer, low back pain, hypertension, and some others.[368] [369] [370] [371] [372] Adding more BQMA privacy protection will reverse people's defensive behaviors and reduce linkage errors, improving data quality, which may improve disease management-enhanced health. The smaller insurer will feel a proportionally smaller health improvement impact as it has fewer policy-holders.

2.4 Chapter Conclusion

A closer analysis of the regulatory, financial, and quality of care data suggests there are benefits to insurance organization from implementing stronger BQMA privacy protections. There is less chance new privacy laws will be passed, burdening the insurer with new compliance requirements. The insurer may experience a positive cash flow because policy-holders are behaving less "defensively" and the insurer has fixed the errors in copy data store(s) identifiers. Disease management protocols may now reduce the insurer's expenses. Better disease management administration will also improve policy-holders' quality of care because policy-holders are better targeted by the disease management protocols. All these results support the insurer's original environmental, economic, and organizational aims, demonstrating the value of the BQMA privacy protections to the insurer and encouraging the adoption of such protections. In the next chapter we explore how to technically accomplish the stronger BQMA privacy protections, providing the insurer the tools it can use.

[368] See Welch, 356.

[369] CorSolutrions, "Cancer Solutions," <http://www.corsolutions.com/programs/2.3.7_cancer.html> (27 October 2005).

[370] Health Management Corporation, "Healthy Returns Program for Low Back Pain," <http://www.choosehmc.com/LowBackPain.html> (26 October 2005).

[371] LifeMasters, "Products & Services," <http://www.lifemasters.com/corporate/prod/index.asp> (26 October 2005).

[372] Accordant, "Rheumatoid Arthritis," <http://www.accordant.net/ra.html> (26 October 2005).

3 Providing Data Privacy

3.1 Applying the Safe Harbor Principle

This thesis will rely on data deidentification to provide application privacy protection. As per the Ethical Force Program's recommendations regarding privacy practices, and HIPAA itself, if data are deidentified consumer privacy concerns should not apply. Data subjects could not be identified for harm. From a software engineering perspective it may be possible to deidentify the BQMA yet permit the applications to operate. This thesis will present a solution for creating additional BQMA privacy protections while preserving sufficient data structure for BQMA and similar applications' operations. The solution includes obtaining computable, linkable results, including some error handling in linkage identifiers. We will work with PM but extend the results to the other BQMA. Since we can't delete all claims data, we must deidentify them. We will work with the UB92 claim record, used by PM. The UB92 has the following key fields:[373] [374] [375] [376]

1) Provider Name/Address/Phone Number
2) Patient Control Number
3) Type of Bill
4) Federal Tax Number
5) Statement Covers Period "From" Date
6) Statement Covers Period "To" Date
7) Billing Covered Days
8) Billing Noncovered Days
9) Coinsurance Days
10) Lifetime Reserve Days
11) Patient Name
12) Patient Address
13) Patient Birth Date
14) Patient Sex
15) Patient Marital Status
16) Admission Date

[373] Centers for Medicare and Medicaid Services, UB92.

[374] See description of UB92 terms in Centers for Medicare and Medicaid services, "Billing Procedures," *Hospital Manual*, 16 September 2004, <http://www.cms.hhs.gov/manuals/10_hospital/ho460.asp> (10 September 2005).

[375] See UB92 description in State of California, Medi-cal, "UB-92 Completion: Inpatient Services," September 2003, <http://files.medi-cal.ca.gov/pubsdoco/publications/masters-MTP/Part2/ubcompip_i00.doc> (10 September 2005).

[376] See UB92 information in Wisconsin Department of Health and Family Services, "UB-92 (CMS 1450) Claim Form Instructions for Personal Care Services," <http://dhfs.wisconsin.gov/medicaid3/updates/2003/2003pdfs/2003-69att4.pdf> (10 September 2005).

17) Admission Hour
18) Admission Type
19) Admission Source
20) Discharge Hour
21) Patient Status
22) Medical Record Number
23) Condition Codes (up to 7 of them)
24) Occurrence Codes (up to 4 of them)
25) Occurrence Dates (up to 4 of them)
26) Occurrence Span "From" Date
27) Occurrence Span "Through" Date
28) Original Document Control Number
29) Value Codes "Code" (up to 3 of them)
30) Value Codes Amount (up to 3 of them)
31) Revenue Code
32) Revenue Description
33) HCPCS/Rates
34) Service Date
35) Service Units
36) Total Charges
37) Noncovered Charges
38) Payer
39) Claim Release Information
40) Provider Number
41) Prior Payments
42) Estimated Amount Due
43) Insured Name
44) Patient Relationship To Insured
45) Patient Identification Number
46) Group Name
47) Insurance Group Number
48) Treatment Authorization Codes
49) Employment Status Code
50) Employer Name
51) Employer Location
52) Principal Diagnosis Code
53) Other Diagnosis Codes (up to 8 of them)
54) Admission Diagnosis Code
55) Principal Procedure Code
56) Principal Procedure Date
57) Other Procedure Codes (up to 5 of them)
58) Other Procedure Dates (up to 5 of them)
59) Attending Physician Id
60) Other Physician Ids (up to 2 of them)
61) Remarks
62) Provider Representative's Signature

63) Provider Representative's Signature Date

This thesis will rely on HIPAA's deidentification standard as the method to provide data deidentification. HIPAA offers two ways to deidentify data. One is for a professional with statistical and scientific knowledge to determine that there is a very small risk that an anticipated recipient of the data can identify the subjects associated with the data.[377] [378] The other method is to use the Safe Harbor principle. Safe Harbor prescribes removing a specific set of items from the data and ensuring that the data producer has no "actual knowledge" that the remaining information can be used alone or in combination with other data to identify the data subjects.[379] In this thesis we will work with the Safe Harbor method. It's easier to follow as its instructions are explicit.

With respect to the "actual knowledge" part of Safe Harbor, several legal experts explained how such a directive can be interpreted. The producer of the deidentified data and his covered entity peers should apply reasonable effort to ensure that data are deidentified based on the potential data recipient.[380] [381] Given the potential sophistication of the recipient and the capability of the covered entity, the covered entity should remove further data to ensure that they cannot be reidentified. Larger covered entities may apply more sophistication. Smaller covered entities, with less knowledge of reidentification techniques, presumably have less deidentification sophistication.[382] More effort will have to be made if the data should be put into the public domain, e.g., online. Users knowledgeable in reidentification—such as that finding combinations of unique items in the data can make some data subjects reidentifiable in certain geographies—will have ready access to this domain and may more readily reidentify data subjects.[383]

Still HIPAA is a new law. Therefore, it's not clear exactly how "actual knowledge" will be interpreted. It's possible that actual knowledge will be interpreted directly as either having such knowledge or not.[384] Safe Harbor may be emphasizing factual knowledge,

[377] HHS (update), 53232.

[378] US Department of Health and Human Services, Office for Civil Rights, "Standards for Privacy of Individually Identifiable Health Information," August 2003, <http://www.hhs.gov/ocr/combinedregtext.pdf> (20 May 2005).

[379] HHS (update), 53232.

[380] Benjamin Butler, attorney specializing in HIPAA, telephone interview with author, May 25, 2005.

[381] Peter Zahn, attorney specializing in HIPAA, telephone interview with author, May 24, 2005.

[382] In several conversations with attorneys for this thesis, the notion of "reasonableness" regarding the deidentification effort was common. (Francesca Brotman-Orner, attorney specializing in HIPAA, telephone interview with author, May 25, 2005; US Department of Health and Human Services, Office for Civil Rights staff, telephone interview with author, February 27, 2004). We adopt such an approach when we conduct an experiment deidentifying data which we obtain from an institution for testing PM deidentification later on in this thesis. We examine the "reasonable" effort that institution may have to undertake to deidentify the data.

[383] For example, see Latanya Sweeney. *Computational Disclosure Control: A Primer on Data Privacy Protection* (Cambridge, MA: Massachusetts Institute of Technology, 2001), 63-82.

[384] Implied Chris Raphaely, attorney specializing in HIPAA, telephone interview with author, April 27, 2005.

intending to give the covered entity more certainty regarding deidentification.[385] Safe Harbor prescribes the following set of items to be removed to deidentify the data set:[386]

1) "Obvious" [identifiers] like name and social security number;
2) all geographic subdivisions smaller than a state, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census: the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and [t]he initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;
3) all elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4) voice telephone numbers;
5) fax telephone numbers;
6) electronic mail addresses;
7) medical record numbers, health plan beneficiary numbers, or other health plan account numbers;
8) certificate/license numbers;
9) vehicle identifiers and serial numbers, including license plate numbers;
10) device identifiers and serial numbers;
11) Internet Protocol address numbers and Universal Resource Locators;
12) biometric identifiers, including finger and voice prints;
13) full face photographic images and any comparable images; and
14) any other unique identifying number, characteristic, or code.

We make several assumptions when deidentifying the UB92. First, the UB92 PHI that will need to be deidentified will refer to patient, Insured, and patient providers' data. The PHI of all three is on the UB92 and must be removed to deidentify such people. As a clarification, the Insured may be paying for the patient's care and may be different from the patient.

We also try to retain some data structure permitting PM operations. A number of items have to be removed per Safe Harbor. PM needs data to identify chronically ill individuals, as before. To understand how to legally leave data in, I conducted interviews with individuals working for PM vendor organizations and insurance companies.[387] I

[385] HHS (part 2), 82542-3.

[386] Taken from University of Miami, "De-identified Health Information (HIPAA)," *Privacy/Data Protection Project*, <http://privacy.med.miami.edu/glossary/xd_deidentified_health_info.htm> (13 October 2003).

[387] Landacorp staff, telephone interview with author, October 1, 2003 and October 16, 2003; SAS technical staff, telephone interview with author, October 2, 2003; Pacificare, Medical Informatics staff, telephone interview with author, October 13, 2003; MEDAI staff, telephone interview with author, October 8, 2003; Medical Scientists technical staff, telephone interview with author, August 22, 2003.

inquired what fields a PM application needs for operations so that a legally-proper deidentified data set can still be created. Based on the interviews, PM only needs certain fields for operations. PM apparently uses only more explicit demographic and medical content. We will conduct an experiment with PM later in this thesis to understand how Safe Harbor obfuscations might impact PM precision; how the application's ability to identify high-risk individuals might change if the data it uses may be removed due to the deidentification. We will leave in some legally-permissible data to permit basic PM computations.

3.1.1 Deidentifying the UB92 Claim Record

Safe Harbor item 7 above suggests that member ids cannot be used. Several such ids exist in the UB92, such as the Patient Control Number, Federal Tax Number, and Provider Number. If a consistent identifier to link policy-holder records can be provided, such explicit ids could be removed.

Safe Harbor items 1, 2, 4, and 14 above suggest that names, phone numbers, or other "unique" type of codes cannot be left in the data. The following are some of such codes:[388]

- Patient Name
- Original Document Control Number
- Insured Name
- Treatment Authorization Codes
- Employer Name

We will remove the non-medical non-demographic codes.

Per Safe Harbor item 3, date of birth (DOB) cannot be available except for year of birth. Further, the PHI of individuals 90 and over should be combined into a category such as "90+". The UB92 has one DOB, Patient Birth Date, which can be modified as needed.[389] [390]

[388] Note, employer information must be removed per the US Department of Health and Human Services, Office for Civil Rights as such information might make some individuals reidentifiable. (Implied, see US Department of Health and Human Services, Office for Civil Rights, "Standards for Privacy of Individually Identifiable Health Information").

[389] For example, DxCG is a popular vendor of PM products used in insurance organizations as we will see later in the text. The DOB appears preferred for DxCG PM platforms, but age, which can be readily obtained through computations using the Safe Harbor-permitted year-of-birth, can also be used. (DxCG, "Technical Requirements," <http://www.dxcg.com/uses/index.html> (22 May 2005)).

[390] Also, an age, as opposed to a date-of-birth variable has been a predictor in various other PM-like platforms, suggesting that age, instead of just DOB, can also be used. (See Symmetry, "A Comparative Analysis of Claims-based Methods of Health Risk Assessment for Commercial Populations," 24 May 2002, 5-6, <http://www.symmetry-health.com/SOAStudy.pdf> (1 September 2005)).

Similarly, Safe Harbor item 3 above forbids other dates (except for year) and dates of activity to be used. Several dates and dates of activities are available in the UB92, such as Admission Date, Occurrence Span "From" Date, and Occurrence Span "Through" Date. These obviously specify the dates and timeframes of the activities associated with data subjects. The dates of activities, however, are typically only used to calculate the length of the activities. All the dates should be given in an annual format in some part of the copy data store, e.g., a new field.[391] For dates that have a "From" and "To" part, the length of the associated activity can be provided, for example, in another field.[392] Subsequently, the dates and dates of activities can be removed as needed information should still be available.

Safe Harbor item 2 forbids the full zip code to be used. There may be three zip codes in the data, within the Patient Address, Provider Name/Address/Phone Number, and Employer Location fields. A deidentified zip code, as described in Safe Harbor item 2, can be used for PM operations. We deidentify the zip code as needed.[393]

This completes Safe Harbor deidentification as the other items in the Safe Harbor list, device identifiers, emails, or biometric identifiers are not in the UB92.

If the producer of the data, e.g., an analyst or his covered entity peers, has knowledge that the resulting UB92-based data set can be used alone or in conjunction with other data to identify the data subjects, the data set can be modified. Based on such "actual knowledge," the characteristics of the people who can be so successfully reidentified can be generalized.[394] Their entire claims records can be deleted. The resulting data should be more protective. We examine the "actual knowledge" aspect of deidentification in our PM experiment later on.

A reidentification mechanism would have to be available. For example, in the PM context, disease management staff would have to initiate contact with PM-uncovered individuals to offer disease management services. Safe Harbor permits the assignment of a "reidentification code" in the data for subsequent re-identification. We incorporate such a code into our deidentification approach. Since we are already deidentifying the member identifiers, to save costs, we can use the replacement of such member identifiers as our reidentification codes. Instead of obfuscating other variables and providing the means to reidentify them, the deidentified member identifiers can serve such purpose.

[391] This is implementation specific and depends on the insurer's current configuration and data management of the copy data store.

[392] This is again implementation specific, depending on the current data management of the copy data store.

[393] When the organization is doing deidentification it can check current Census data to find which current 3-digit zip codes need to become 000. Based on Census 2000 data, HIPAA provides the 17 zip codes which must be changed to 000 because their associated geographic units contain no more than 20,000 people. (HHS (update), 53234). The 3-digit zip codes are: 036, 059, 063, 102, 203, 556, 692, 790, 821, 823, 830, 831, 878, 879, 884, 890, and 893. We will assume these are the zip codes which need to be changed in 2005 and 2006 as well for illustration when we conduct our PM experiment later on.

[394] For example, see Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection,* 63-82.

Note, HIPAA's Safe Harbor principle states that the "reidentification code" cannot be "derived" from or related to information about the data subject.[395] It would seem that transforming linkage identifiers into secure reidentifiable identifiers, as will be suggested and explored by this thesis, should not be allowed. Based on other information such a conclusion might not necessarily be the case. If there is a secure way to transform PHI into a reidentification code, this should be permitted. The spirit of HIPAA is to protect data. If a strong mathematical function can be used to generate a reidentification code, such a function should be permissible. Indeed, one can look at this question from a statistical point of view. HIPAA allows a statistician to render data deidentified. If a statistician feels the deidentified data have a low risk of being reidentified, they are considered secure. Safe Harbor as a method should already presumably be low-risk because it was designed so that information subject to its tenets could be used for any purpose, including being placed in the public domain.[396] If any transformation function creating the reidentification code is shown to provide low risk it should be acceptable with regard to the statistician's method. The transformation creates "low risk" data. We will attempt such a secure transformation in this thesis.[397]

Finally, as there are errors in the member identifiers, the reidentification code will have to handle such errors. Resources might not be available to clean the errors earlier.[398]

[395] See discussion in HHS (update), 53232.

[396] Quintiles.com, "Understanding the Impact of HIPAA on Clinical Research," June 26-27, 2003. <http://www.quintiles.com/NR/rdonlyres/e7n4reuv4qqzqtjpdpzirnsyzbhmygzs4uijtkagmy5gie5dvu5gvyhkmme5dwrbbcmoffkiiteejc/JBeachBarnett03.pdf> (31 August 2005).

[397] HIPAA also states that the reidentification code cannot be "used" or "disclosed" for any purpose except reidentification. (See discussion in HHS (update), 53233). The use of any such code for any data linkage, as will also be examined in this thesis, should be prohibited, too. This may not necessarily be the case either. For example, a considerable amount of epidemiological research relies on record linkage of health data. (William Winkler, "The State of Record Linkage and Current Research Problems," <http://www.census.gov/srd/papers/pdf/rr99-04.pdf> (12 March 2004)). Such linkage can be based on a consistent code such as a reidentification code. (Fritz Scheuren and William Winkler, "Regression Analysis of Data Files That Are Computer Matched – Part I," *National Research Council. Record Linkage Techniques – 1997: Proceedings of an International Workshop and Exposition* (Washington, DC: National Academy Press, 1999), 106). The latter may be a consistent code if it's created consistently for all equal identifiers. Indeed, when used with deidentified data, linking records in and of itself may not necessarily facilitate the reidentification of data subjects. For example, in the BQMA context, the intent of linking records is to better manage the organization, better manage groups of policy-holders with chronic conditions, etc., not to reidentify individuals per se for any direct harm. The intent of HIPAA's confidentiality should be preserved as no reidentification or direct harm should take place. Also, as mentioned before, HIPAA designers encouraged the use of deidentified data for purposes when this was possible. The creation of a reidentification code which can be used for linking records helps in the creation of deidentified data. Data can be linked and analyzed. Finally, this thesis actually proposes to create several such secure reidentification codes. All will be based on linkage identifiers such as medical record numbers. HIPAA only discusses the assignment of one code to a deidentified data set. However, the creation of two or more codes, as long they meet HIPAA's requirements, should be acceptable. (Brian Annulis, attorney specializing in HIPAA, telephone interview with author, May 3, 2005).

[398] A solution to securely link records for applications that link data should handle errors in the linkage identifiers. Ideally, errors would be cleaned before the application uses the data so that the application does not deal with errors. For the BQMA, it might not be possible to remove the errors in the identifiers beforehand. Presumably errors in the linkage identifiers happen upstream before data reach the main linkage data set. There might be more than one data source that feeds the copy data store. (Health Plan Employer Data and Information Set (HEDIS) auditor, telephone interview with author, May 19, 2004).

The table below illustrates the changes to a hypothetical UB92 after deidentification by Safe Harbor. Any further data removal based on "actual knowledge" will be explored later on in our PM experiment.

Therefore, any cleaning solutions may have to be placed at more than one upstream location, probably increasing installation costs. It may not be easy placing such solutions upstream; the processes or people responsible for placing data into the linkage data set may not have the proper authority or the storage capacity required. (Implied, Dorothy Curtis, Research Scientist, Lab for Computer Science, MIT, personal interview, January 20, 2004). Their focus may not include data linkage for applications like the BQMA. They may not be able to access all the data all the time or have the storage required to store any redundant data necessary for error resolution (potentially, stemming from the same data access restrictions, etc). The need to handle errors rather than cleaning them upstream is even more relevant for data linkage applications outside of an organization. Here the file to be linked is transmitted to another entity, for example, an epidemiologist, for analysis. (See discussions in Shaun Grannis, J. Marc Overhage, and Clement McDonald, "Real World Performance of Approximate String Comparators for Use in Patient Matching," 2004, 43. <http://www.cs.mun.ca/~harold/Courses/Old/CS6772.F04/Diary/5604Grannis.pdf> (22 May 2005); L. L. Roos and A. Wajda, "Record Linkage Strategies: Part I: Estimating Information and Evaluating Approaches," *Methods of Information in Medicine*, 30 (1991): 117-118). It would be difficult if not impossible for a receiving organization to clean any errors *upstream*. Presumably, the receiving entity has little influence over the original data-producing organization as it does not manage it. Therefore it might not be able to place any cleaning solutions upstream. The field of record linkage we discuss later in the text exists because data-producing organizations may not clean errors, yet records must be linked. The receiving entities must deal with errors. Our approach must deal with such errors, too.

UB92 Claims Data Before and After Deidentification

Before Deidentification	**After Deidentification**
1) Provider Name/Address/Phone Number	1) *Delete. Except Zip Code (Should be 3 Digits or "000" When Pop. ≤ 20,000)*
2) Patient Control Number	2) *delete*
3) Type of Bill	3) Type of Bill
4) Federal Tax Number	4) *delete*
5) Statement Covers Period "From" Date	5) *Provide in Annual Format Elsewhere*
6) Statement Covers Period "To" Date	6) *Provide Length of "Statement Period" ("To" date - "From" date) Elsewhere*
7) Billing Covered Days	7) Billing Covered Days
8) Billing Noncovered Days	8) Billing Noncovered Days
9) Coinsurance Days	9) Coinsurance Days
10) Lifetime Reserve Days	10) Lifetime Reserve Days
11) Patient Name	11) *delete*
12) Patient Address	12) *Delete. Except Zip Code (Should be 3 Digits or "000" When Pop. ≤ 20,000)*
13) Patient Birth Date	13) *Provide Age; Also Use 90+*
14) Patient Sex	14) Patient Sex
15) Patient Marital Status	15) Patient Marital Status
16) Admission Date	16) *Provide in Annual Format Elsewhere*
17) Admission Hour	17) Admission Hour
18) Admission Type	18) Admission Type
19) Admission Source	19) Admission Source
20) Discharge Hour	20) Discharge Hour
21) Patient Status	21) Patient Status
22) Medical Record Number	22) ***Encrypt Medical Record Number***
23) Condition Codes (up to 7 of them)	23) Condition Codes (up to 7 of them)
24) Occurrence Codes (up to 4 of them)	24) Occurrence Codes (up to 4 of them)
25) Occurrence Dates (up to 4 of them)	25) *Provide in Annual Format Elsewhere (up to 4 of them)*
26) Occurrence Span "From" Date	26) *Provide in Annual Format Elsewhere*
27) Occurrence Span "Through" Date	27) *Provide Length of Stay ("Through" date - "From" date) Elsewhere*
28) Original Document Control Number	28) *delete*
29) Value Codes "Code" (up to 3 of them)	29) Value Codes "Code" (up to 3 of them)
30) Value Codes Amount (up to 3 of them)	30) Value Codes Amount (up to 3 of them)
31) Revenue Code	31) Revenue Code
32) Revenue Description	32) Revenue Description
33) HCPCS/Rates	33) HCPCS/Rates
34) Service Date	34) *Provide in Annual Format Elsewhere*
35) Service Units	35) Service Units
36) Total Charges	36) Total Charges
37) Noncovered Charges	37) Noncovered Charges

38) Payer	38) Payer
39) Claim Release Information	39) Claim Release Information
40) Provider Number	40) *delete*
41) Prior Payments	41) Prior Payments
42) Estimated Amount Due	42) Estimated Amount Due
43) Insured Name	43) *delete*
44) Patient Relationship To Insured	44) Patient Relationship To Insured
45) Patient's Identification Number	45) *delete*
46) Group Name	46) Group Name
47) Insurance Group Number	47) Insurance Group Number
48) Treatment Authorization Codes	48) *delete*
49) Employment Status Code	49) Employment Status Code
50) Employer Name	50) *delete*
51) Employer Location	51) *Delete. Except Zip Code (Should be 3 Digits or "000" When Pop. ≤ 20,000)*
52) Principal Diagnosis Code	52) Principal Diagnosis Code
53) Other Diagnosis Codes (up to 8 of them)	53) Other Diagnosis Codes (up to 8 of them)
54) Admission Diagnosis Code	54) Admission Diagnosis Code
55) Principal Procedure Code	55) Principal Procedure Code
56) Principal Procedure Date	56) *Provide in Annual Format Elsewhere*
57) Other Procedure Codes (up to 5 of them)	57) Other Procedure Codes (up to 5 of them)
58) Other Procedure Dates (up to 5 of them)	58) *Provide in Annual Format Elsewhere (up to 5 of them)*
59) Attending Physician Id	59) *delete*
60) Other Physician Ids (up to 2 of them)	60) *delete*
61) Remarks	61) *delete*
62) Provider Representative's Signature	62) *delete*
63) Provider Representative's Signature Date	63) *Provide in Annual Format Elsewhere*

In summary, to satisfy Safe Harbor while making the data initially usable for PM entails the following changes. We call them Enhancements 1 through 6 from now on:

1) Enhancement 1. Provide a consistent id unrelated to other member identifiers in the data. This id should be reidentifiable to obtain an original member id and it should handle errors in the original identifier. All other member identifiers should be removed.
2) Enhancement 2. Remove all other unique values or codes in the data.
3) Enhancement 3. For the data subjects, provide year of birth instead of date of birth, and a "90+" category when individuals are at least 90 years old.
4) Enhancement 4. Provide length of activity instead of dates of activity, and provide the year component only for one of the two dates-of-activity dates as well as all the single dates. Remove all the original date variables.
5) Enhancement 5. Generalize the zip code fields as prescribed by Safe Harbor item 2, as described earlier.
6) Enhancement 6. Apply reasonable effort to remove any other data so there is no "actual knowledge" the data can be used alone or with other data to reidentify the data subjects.

3.2 *Quantifying the Value of Information*

We demonstrate how to technically accomplish Enhancements 1 through 6. They should be solved with data perturbation techniques. We create an encryption and modification method to mask data while permitting PM functionality. We make several assumptions. First, any linkage file involved may be large. According to the Centers for Disease Control and Prevention, in 2003, approximately 85% of individuals with health insurance saw a health professional within the past 12 months at least once.[399] Roughly 80% of these individuals saw a health professional multiple times. In 2002, at least 70% of individuals' total health care expenses were paid by private or public insurance programs.[400] Thus, claims forms to insurance organizations should be generated by the majority of people's visits to providers. An average size insurance organization may have almost 154,000 policy-holders.[401] Assuming several claims get paid for by the insurer for each policy-holder, on average, the copy data store in an average insurer may be filled with a half million new claim records annually from the above computations. This is

[399] Centers for Disease Control and Prevention, "Summary Health Statistics for US Adults: National Health Interview Survey, 2003," July 2005, <http://www.cdc.gov/nchs/data/series/sr_10/sr10_225.pdf> (31 August 31, 2005).

[400] Medical Expenditures Panel Survey, "2002 Compendium of Tables – Household Medical Expenditures," 23 December 2004, <http://www.meps.ahrq.gov/mepsnet/tc/TC15.asp?_SERVICE=MEPSSocket1&_PROGRAM=MEPSPGM.TC.SAS&File=HCFY2002&Table=HCFY2002%5FPLEXP%5F%40&VAR1=AGE&VAR2=SEX&VAR3=RACETHNX&VAR4=INSURCOV&VAR5=POVCAT02&VAR6=MSA&VAR7=REGION&VAR8=HEALTH&VARO1=5+17+44+64&VARO2=1&VARO3=1&VARO4=1&VARO5=1&VARO6=1&VARO7=1&VARO8=1&_Debug> (31 August 2005).

[401] America's Health Insurance Plans, <http://www.ahip.org/> (31 August 2005).

given the just-described submission of claims by policy-holders. In a large insurer, with the number of policy-holders several times the 154,000 figure, the copy data store may be filled with several million new policy-holder claims annually. Another assumption is that the linkage identifiers involved may be diverse—unique within the linkage file. Since most people annually use the health care system, the copy data store should have numerous unique member ids present as claim records for many unique policy-holders should be submitted. Finally, the linkage file may be dense from the point of view of the linkage identifiers. For instance, in the PM context member ids may be alphanumeric.[402] On occasion, member ids are given out in sequential order within insurance organizations.[403] One new policy-holder might get id 80155243 while the next new policy-holder will get id 80155244. In this case, the copy data store should have a number of "close" alphanumerical member ids for any given member id as many unique member ids would exist in the copy data store. Identifiers in linkage files in other applications might be diverse and dense on occasion as well.

3.2.1 Addressing Data Privacy

Data perturbation solutions can be broadly classified as encryption and hashing techniques and non-encryption non-hashing solutions. Encryption can be defined as a method that transforms original information, *plaintext*, into altered information, *ciphertext*, which appears as nonsensical to an observer.[404] A hashing function transforms a plaintext message into a hash, a concise representation of the original, usually longer plaintext.[405] [406] Based on Menezes' taxonomy of cryptographic primitives, encryption and hashing systems may be keyed or unkeyed.[407] A "key" may designate the nature of the transformation between plaintext and ciphertext, facilitate the encryption process, or facilitate the hashing process.[408] Non-encryption non-hashing data perturbation solutions include swapping or suppression.[409] For example, suppressing a value obviously removes the value.[410]

Providing Enhancements 2 through 6 should be done using non-cryptographic techniques. Using cryptographic techniques for such enhancements may not be efficient. Certainly some of the data to be removed, including that which should be removed based

[402] DxCG, *DxCG RiskSmart StandAlone User Guide Version 2.0.1*, January 2005, 24.
[403] Customer Service Staff, Harvard Pilgrim, telephone interview with author, June 9, 2004.
[404] See Deborah Russell and G.T. Gangemi, Sr., "Chapter 3: Computer System Security and Access Controls," *Computer Security Basics*, <http://www.oreilly.com/catalog/csb/chapter/ch03.html> (14 October 2003).
[405] See Alfred Menezes, Paul C. van Oorschot and Scott A. Vanstone, "Chapter 1," *Handbook of Applied Cryptography*, 2001, 5, <http://www.cacr.math.uwaterloo.ca/hac/about/chap1.pdf> (22 May 2005).
[406] RSA Security, "What Is a Hash Function?" <http://www.rsasecurity.com/rsalabs/node.asp?id=2176> (22 May 2005).
[407] Menezes, "Chapter 1," 5.
[408] Menezes, "Chapter 1," 11, 27.
[409] See for example Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection*, 63.
[410] See Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection*, 56.

on "actual knowledge," can be encrypted or hashed. However, adding a more complicated cryptographic process to applications rather than transforming data directly might be costlier. Software may have to be integrated into the application to perform the transformations. Functionality for non-cryptographic techniques should probably already exist within the software platform, as such functions should provide standard modifications to data in almost any format. We recommend a non-cryptographic approach for Enhancements 2 through 6. The deidentification effort involved should be straightforward, performed as detailed in Section 3.1.1.

Conversely, providing Enhancement 1 should be done via an encryption mechanism. A method is needed that obfuscates linkage identifiers, facilitates their linkability, yet handles errors within the identifiers. Available approaches do not appear suitable. This thesis proposes a new encryption-based approach to handle such requirements.

To provide Enhancement 1, however, we switch from the PM world to address the more general world of record linkage. In resolving Enhancement 1 we need a general solution as the just-stated requirements are general.

No current approach appears to handle sufficiently all requirement aspects from above. For example, there exist approaches to hash linkage identifiers, securely evaluate any generic function, and create matrices to hold comparison results of identifiers with errors. They will be examined later on. These approaches may not address all aspects of Enhancement 1. Some approaches may not handle all errors; other approaches might be inefficient; others may not protect all aspects of privacy.

We create a new approach. In our approach we explain how to locate the identifiers that can address the errors of linkage identifiers by obtaining additional data on which records should be linked. We examine linkage fields' "information content," including asking when fields should be analyzed at the character-level to reduce linkage errors. Practitioners can decide whether to analyze a given field at the character-level or in its entirety to overcome their software's linkage identifier errors. We devise a threat model that can be used to judge the privacy protection offered by an approach. This is a generic paradigm for evaluating application security within and outside of an original organization. Since we are focused on the BQMA—internal applications—practitioners can ask if an approach offers appropriate security if the data involved are produced by the same or different organizations. If it is the same organization, more security would be necessary since employees might have access to internal data deidentification processes against which a security approach must protect. Finally, we provide a solution that meets our threat model. We describe its operations and security, and demonstrate its strength over other existing approaches.

3.2.2 Record Linkage of Data

The process for reducing identifier errors in applications linking records is described. In the below, we first describe the purpose of record linkage and how likelihood ratios and linkage thresholds form the basis of distinguishing true from false links. Next we use such constructs in selecting fields within records that will provide optimum linkage. Finally, character-level analysis of fields is explored to show under what conditions exposing a field's characters for analysis might not be useful for linkage due to the characteristics of the linkage data set. This analysis has implications for understanding under what linkage data set characteristics analyzing a field's characters is useful.

3.2.2.1 Record Linkage Concepts

Record linkage is the process of combining two or more records to link information relating to a single unit, such as an individual, a family, or an event.[411] Numerous applications utilize such a technique, often answering questions on relationships or effectiveness of the associated programs, people, or entities. For example, linking police records and court records is useful in answering questions such as which variables (e.g., type of assault, location of a break-in, etc.) affect severity of a prison sentence.[412] Hospital discharge data can be linked to themselves to determine if the length of a newborn's postnatal hospital stay is related to his future hospital readmissions.[413] The hypothesis may be that the shorter the hospital stay the more probable the readmission. When a unique identifier is unavailable for such linkage, however, because the identifier is in a different format, it's not unique, or it has errors, including missing values, more advanced techniques are required. [414] [415] [416] [417]

Our interest is to deidentify record linkage as per Enhancement 1. In this chapter, we therefore ask the following two questions: What deidentification approaches preserve record linkage? What are the best ways of securing data in the record linkage context?

[411] Martha Fair, "Recent Developments at Statistics Canada in the Linking of Complex Health Files," <http://www.fcsm.gov/99papers/fair.pdf> (22 May 2005).

[412] Scott Meyer, "Using Microsoft Access to Perform Exact Record Linkages," 1997, 280, <http://www.fcsm.gov/working-papers/smeyer.pdf> (22 May 2005).

[413] Shiliang Liu and Shi Wu Wen, "Development of Record Linkage of Hospital Discharge Data for the Study of Neonatal Readmission," 2000, <http://www.phac-aspc.gc.ca/publicat/cdic-mcc/20-2/c_e.html> (22 May 2005).

[414] See Winkler, "Preprocessing of Lists and String Comparison," 181.

[415] See Scheuren, 106.

[416] William Winkler, "Matching and Record Linkage," 1997, 378, <http://www.fcsm.gov/working-papers/wwinkler.pdf> (21 May 2005).

[417] Winkler, "Preprocessing of Lists and String Comparison," 183-5.

3.2.2.2 Record Linkage Operations

To conduct our analysis we first describe basic record linkage operations. The purpose of record linkage is to determine which records represent matches and which represent non-matches in a file.[418] [419] Below, a *match* will represent the case when the two records being compared represent the same unit, for example, the same person; a *non-match* will represent the case when the two records represent two different units, e.g., two different individuals. There are two steps involved: training and validation. Training configures the record linkage system for subsequent use. The system then uses the training parameters to run on actual, validated data. In this thesis, we focus mostly on the training step. We will optimize record linkage training parameters to improve validation operations.

We will link a single file to itself, just as in the PM case. However, our analysis will apply to applications that link two (or more) files, as is more common in record linkage. The fundamental computation behind record linkage is that of likelihood ratios, representing the odds that any two records represent the same unit. High ratio values imply a match. Low ratio values imply a non-match, as we will see below. Each training file record is linked with every other. In the resulting cross product, the system must determine if record R_i and record R_j represent a match for all i, j=1…N, where N is the number of records in the file. "Patterns" are used to identify matches. That is, questions about a comparison of the fields making up R_i with the corresponding fields in record R_j are computed for every record pair. Almost any computation is possible.[420] The pattern may be that two corresponding fields must agree exactly on their contents; the pattern may be that several corresponding fields should specifically disagree on their last 3 characters; or the pattern may be that two corresponding fields should be equal to "Jones."

Given observed patterns 1 through k, a likelihood ratio (LR) is formed during training. The computed LR is

$$\text{LR} = \quad \text{P(pattern}_1,\ldots,\text{pattern}_k \mid R_i \text{ and } R_j \text{ match) /} \\ \text{P(pattern}_1,\ldots,\text{pattern}_k \mid R_i \text{ and } R_j \text{ don't match)} \qquad (1)$$

The intent of training is to find patterns such that the LR created would properly designate the match status of R_i and R_j when the system actually observes the pattern during training. The system will distinguish matching from non-matching records by separating high and low LRs. Useful patterns would create LRs significantly different from 1. Otherwise, it would be difficult to distinguish whether the two records match upon observing the pattern. Equality, inequality, or partial equality of fields are commonly used patterns because they will generate LRs significantly higher and lower

[418] For example, see Martha Fair, "Recent Developments at Statistics Canada in the Linking of Complex Health Files."
[419] Winkler, "The State of Record Linkage and Current Research Problems."
[420] See Ivan Fellegi and A.B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64 (1969): 1185.

than 1, respectively, allowing for the separation of matches versus non-matches, as we will see below.

Since the possible number of patterns involving all record fields is vast, one common assumption made in record linkage is that of conditional independence. The comparisons represented by patterns are assumed to be mutually independent conditioned on either records matching or not matching. For example, imagine two patterns: equality of the date of birth (DOB) field and equality of the street name field. Whether two records agree on DOB is assumed to be independent of whether they agree on the street name of the address if we know that the records match or don't match. Practically speaking, in real life, we don't anticipate the DOB to be "correlated" to street name. The ideas we describe in this thesis, however, will hold even if this assumption is not true, and certain fields are correlated. Equation (1) can be rewritten as:

$$\begin{aligned} \text{LR} = \quad & \text{P(pattern}_1 \mid \text{R}_i \text{ and R}_j \text{ match) * ... * P(pattern}_k \mid \text{R}_i \text{ and R}_j \text{ match) /} \\ & \text{P(pattern}_1 \mid \text{R}_i \text{ and R}_j \text{ don't match) * ... * P(pattern}_k \mid \text{R}_i \text{ and R}_j \text{ don't match) =} \\ & \\ & \text{[P(pattern}_1 \mid \text{R}_i \text{ and R}_j \text{ match) / P(pattern}_1 \mid \text{R}_i \text{ and R}_j \text{ don't match] *... *} \\ & \text{[P(pattern}_k \mid \text{R}_i \text{ and R}_j \text{ match) / P(pattern}_k \mid \text{R}_i \text{ and R}_j \text{ don't match]} \qquad (2) \end{aligned}$$

For ease of analysis, it is customary to work with the log of the LR computed above.[421] We have

$$\begin{aligned} \text{log (LR)} = {} & \text{log (P(pattern}_1 \mid \text{R}_i \text{ and R}_j \text{ match) / P(pattern}_1 \mid \text{R}_i \text{ and R}_j \text{ don't match)) + ... +} \\ & \text{log (P(pattern}_k \mid \text{R}_i \text{ and R}_j \text{ match) / P(pattern}_k \mid \text{R}_i \text{ and R}_j \text{ don't match))} \qquad (3) \end{aligned}$$

Each of such k terms is considered a weight. We refer to the above computed log (LR) as $W_{\text{total weight}}$ from now on as it sums the k weight terms.

There are several ways to select the patterns and compute each of the k weight terms for record linkage. Examining the training file directly, using weights from prior linkages, and theoretically determining what should be the weights in a typical distribution of units are some of the techniques available.[422] [423] One approach is to analyze the data directly. When analyzing a training file, for example, the match status of all or a representative sample of records in the cross product would be examined. For this sample, the true match status regarding any two records should be obtained. Patterns and weights can be set: patterns should be found to force weights for matching records when patterns agree to be higher. Weights for non-matching records in the sample should be lower, and patterns should be found so that non-matching units should disagree. Matching will improve. If high weights indicate matches and low weights non-matches, the above construction will allow for better identification of matches versus non-matches.

[421] Tony Blakely and Clare Salmond, "Probabilistic Record Linkage and a Method to Calculate the Positive Predictive Value," *International Journal of Epidemiology*, 31 (2002): 1247.

[422] For example, see Martha Fair and Patricia Whitridge, "Tutorial on Record Linkage Slides Presentation," 1997, 463, <http://www.fcsm.gov/working-papers/mfair-tutorial.pdf> (23 May 2005).

[423] Blakely, 1247.

That is, during matching $W_{total\ weight}$, comprised of all the weights, is compared to a threshold. If $W_{total\ weight}$ equals or exceeds the threshold, the system will consider the two records a match. If $W_{total\ weight}$ falls below the threshold, the pair will be deemed a non-match. In more sophisticated record linkage approaches two thresholds may be used.[424] [425] When $W_{total\ weight} \geq$ Upper Threshold the record pair is deemed a match. When $W_{total\ weight} \leq$ Lower Threshold the record pair is considered a non-match. For simplicity of exposition, we will not be using the dual-threshold model. Our results are generalizable to a dual-threshold model, however.

The threshold should be set to maximize the usefulness of the linkage. Four outcomes are possible when determining match status of two file records: true positive, false positive, true negative, and false negative. The false negatives and false positives arise due to errors or the natural distribution of field values. For example, two different people can have the same date of birth (DOB). If DOB is a comparison field and other comparison fields for these two individuals do not produce appropriate weights low enough, the equality of the DOB may create a false positive if the DOB weight is high enough to make the $W_{total\ weight}$ higher than the threshold. The two different people will be falsely labeled a match. A utility could be assigned to each of the four outcomes above and a threshold be computed which maximizes total system utility.[426] Practically, the threshold is often set by examining the cross product in the training file and assessing the type of errors that can be tolerated by the application. If the intent is to avoid false negatives, the threshold should be set to a low $W_{total\ weight}$. Matching records should have weights above that cutoff as patterns were chosen to create higher weights.[427] If the intent is to have no false positives, the threshold should be set high. Weights for non-matching records should be lower than that threshold, as patterns were chosen to create lower weights for non-matching records. In our single-threshold system we will arbitrarily define the threshold as halfway between the maximum and minimum weights for compared fields. Our approach will generate some false positives and some false negatives. The total weights of some matching units will fall below the threshold, and the weights of some non-matching units will appear as matches. We use a simpler threshold for illustration. Our analysis can be undertaken using other choices of thresholds.

Further, it may be possible to attribute greater "importance" to different weights and create linkage field *coefficients* which would more optimally interact with a given threshold. Field weights could be multiplied by field coefficients signifying "importance" to indicate the usefulness of a particular field. However, the patterns themselves may be constructed to create appropriate field weights to optimize field relationships to

[424] Winkler, "Matching and Record Linkage," 382-3.

[425] Shaun Grannis, J.M. Overhage, S. Hui, and C.J. McDonald, "Analysis of a Probabilistic Record Linkage Technique Without Human Review" (American Medical Informatics Association 2003 Symposium Proceedings), 259.

[426] For setting utilities in record linkage applications see similar ideas in C. Quantin, C. Binquet, F.A. Allaert, B. Cornet, R. Pattisina, G. Leteuff, C. Ferdynus, and J.B. Gouyon, "Decision Analysis for the Assessment of a Record Linkage Procedure," *Methods of Information in Medicine*, 44 (2005): 77.

[427] For instance, see David White, "A Review of the Statistics of Record Linkage for Genealogical Research," 1997, 368-71, <http://www.fcsm.gov/working-papers/dwhite.pdf> (23 May 2005).

thresholds. Since weights are constructed based on patterns, the pattern itself can be chosen to create the needed weight "importance." Importance can be specified by creating multiple similar patterns, patterns based on individual characters analogous to patterns encompassing full fields, and similar constructions, creating multiple weights for a field. The diversity of weights, instead of using fixed coefficients, could offer more flexibility in constructing field weights to optimize matching.

Once the system is trained it can be run on validated data. The weights and threshold computed above are used for matching. The same process is followed as above with respect to computing $W_{\text{total weight}}$ and comparing it to a threshold to determine match status of a validation record pair in a validation file.

3.2.2.3 Field Information Content

3.2.2.3.1 Selecting Fields for Linkage

We work with the basic patterns of full and partial field equality and inequality to develop our theory for minimizing the impact of identifier errors. Assuming an error rate for linkage identifiers which is not significant, as per the discussions in Sections 1.3.2 and 2.2.5, a single weight computed as

$\log(P(\text{pattern}_q \mid R_i \text{ and } R_j \text{ match}) / P(\text{pattern}_q \mid R_i \text{ and } R_j \text{ don't match}))$

will be significantly higher than 1 when R_i and R_j match and pattern_q is the equality pattern. The same units should typically have equal identifiers, creating a high numerator above. Similarly, the above weight will be significantly lower than 1 when R_i and R_j don't match and pattern_q is the inequality pattern. Typically when units are different their identifiers will not equal, creating a much higher denominator above. High and low weights can now be separated.

Some record linkage practitioners recommend using more fields to reduce errors from suboptimal identifiers.[428] Others recommend employing string comparators—which examine the characters of compared variables and assign a similarity "score" based on equality of certain characters—to reduce errors.[429] [430] [431] More information can be used to add certainty in overcoming linkage problems. How do we know which fields in records have the best "information?" How do we know which to analyze at the character level? Current literature discusses reasons why certain variables appear more useful than

[428] For instance, see Winkler, "Matching and Record Linkage," 393.

[429] Winkler, "Preprocessing of Lists and String Comparison," 185.

[430] See Fred Damerau, "A Technique for Computer Detection and Correction of Spelling Errors," *Communications of the ACM,* 7 (1964): 172.

[431] Grannis, "Real World Performance of Approximate String Comparators for Use in Patient Matching," 46.

others.[432] [433] [434] Some literature talks about the "information content" of a variable and asks how beneficial it might be for linkage.[435] [436] However, current literature doesn't explain the tradeoffs between the errors of a field and linkage outcomes. The authors may resolve errors in particular systems without describing more generally how to improve error handling in a variety of contexts. It may be less clear how to overcome a field's errors in other linkage projects, which may or may not have similar traits to projects already examined.

This thesis explores such questions. It examines how a field's errors might impact linkage. We explore the usefulness of character-level analysis, describing the benefit of the *new* information. This thesis will not create a complete analytical framework. Future research can create a more generic approach.

We first ask how to select fields to add linkage redundancy to address linkage identifier errors. Imagine records R_i and R_j represent the same unit. However, the comparison fields involved have sufficient error such that their computed $W_{total\ weight}$ is below the threshold. The system would incorrectly label R_i and R_j a non-match. If we can find another field such that its weight pushes $W_{total\ weight}$ above the threshold, the records will properly be designated as a match. Similarly, to fix a false positive, we must find another field whose disagreement weight pulls $W_{total\ weight}$ below the threshold to rectify this error.

3.2.2.3.2 Using a Single Linkage Field

We will work with the following example to demonstrate our analysis. Imagine that an existing system uses only a field arbitrarily named K_1. Assume the number of character positions in K_1 is n1, and each character position has a range of p_1 values. For example, p_1 would be 10 if each K_1 character is a digit ranging 0-9. K_1 has an error rate e, in [0,1]. With probability e, K_1 is subject to a typo, insertion, missing value or similar errors that can happen to a field. We assume the K_1 namespace, p_1^{n1}, is large and the error rate is small.[437] Consider field K_2, another field available in the records. The number of character positions in K_2 is n2 and each K_2 character has a range of p_2 values. Assume the

[432] Catherine Quantin, C. Binquet, K. Bourquard, R. Pattisina, B. Gouyon-Cornet, C. Ferdynus, J.B. Gouyon, and F.A. Allaert, "A Peculiar Aspect of Patients' Safety: The Discriminating Power of Identifiers for Record Linkage," *Studies in Health Technology and Informatics*, 103 (2004): 400-406.

[433] Quantin, "Decision Analysis for the Assessment of a Record Linkage Procedure," 72-9.

[434] Shaun Grannis, J.M. Overhage, and C.J. McDonald. "Analysis of Identifier Performance Using a Deterministic Linkage Algorithm" (Proceedings of the AMIA 202 Annual Symposium), 305-9.

[435] L. J. Cook, L.M. Olson, and J.M. Dean. "Probabilistic Records Linkage: Relationships Between File Sizes, Identifiers, and Match Weights," *Methods of Information in Medicine*, 40 (2001): 196-203.

[436] Roos, 117-23.

[437] Obviously not all namespaces will be large. Our analysis will work with smaller namespaces, but we use a larger namespace for illustration. Our analysis will also work with larger errors, but again we use a small error rate for presentation. Nevertheless, a larger namespace and smaller error rate are realistic for a number of linkage applications. For example, the Patient Account Number may be a large field, with an arbitrary length of 10 positions, for example, within a health organization. If it is the primary identifier for linking patient data it may have a low error rate given its importance and the quality monitoring to which it probably is subject.

K_2 namespace, p_2^{n2}, is also large. K_2 has error rate f. The same types of mistakes can happen to K_2 as for K_1. Likewise, assume a small K_2 error rate. We assume a uniform distribution of all values in the namespaces for K_1 and K_2 for simpler presentation although our approach will work with more complex namespace distributions. We'd like to understand what characteristics K_2 must meet to fix the errors created by a system only using K_1. Section 1.3.2 described how errors in the medical record number can lead to poor linkages. We will minimize them in this presentation by using redundant data. The Levenshtein string comparator will be used in our analysis. The Levenshtein string comparator measures the distance between two strings by determining the smallest number of insertions, deletions, and substitutions needed to change one string into another.[438]

Imagine that the current system compares equality and inequality of full K_1 values *without* any character-level analysis of K_1. We compute the weights for the system to understand system operations. The weights will be determined theoretically. By computing the likelihood ratios, we see that for this system the *agreement* weight, representing the equality pattern, and *disagreement* weight, representing the inequality pattern are:

$$[K_1]W_{\text{agree}} = \log\left(P(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) / P(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})\right) \qquad (4)$$

$$[K_1]W_{\text{disagree}} = \log\left(P(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) / P(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})\right) \qquad (5)$$

Since the overall error rate for K_1 is e, we have:
$P(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) = e*(1-e) + e*(1-e) + e^2 - (e^2) * (1/p_1)^{n1}$

The first term on the right represents $R_i[K_1]$ being in error when $R_j[K_1]$ is not in error. The second term represents the opposite event. The third term represents the event when the K_1 fields of both records are in error. The last term computes the chance that both fields are in error and, uncommonly, the errors produce identical K_1 values. The possibility of two K_1 fields randomly equaling is one chance within the size of the namespace, represented by $(1/p_1)^{n1}$. The full fourth term must be subtracted from the third term, e^2, since this event quantifies the K_1 values of both fields erroneously equaling rather than not as the conditional probability above requires. We can drop all the terms which contain e^2 terms as e is assumed small. We have
$P(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) = e - e^2 + e - e^2 + e^2 - (e^2) * (1/p_1)^{n1}$
$P(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) \approx 2*e \qquad (6)$

Consequently,
$P(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) \approx 1 - 2*e \qquad (7)$

On the other hand, we have,

[438] Grannis, "Real World Performance of Approximate String Comparators for Use in Patient Matching," 44.

$P\ (R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match}) =$
$[2 * e * (1\text{-}e) * (1/p_1)^{n1}] + ((1\text{-}e)^2) * (1/p_1)^{n1} + (e^2) * (1/p_1)^{n1}$

The rightmost $(1/p_1)^{n1}$ term represents the possibility of the K_1 fields of two non-matching records equaling when both K_1 fields are in error. This is the size of the K_1 namespace: one value in each character position in the field randomly being chosen across all characters positions independently. The associated e^2 term indicates a mistake took place in both fields for both of them to equal erroneously. The middle $(1/p_1)^{n1}$ term represents the possibility of K_1 fields of two non-matching records equaling due to the natural distribution of field values, which is again the size of the K_1 namespace: one value in each field character position randomly being chosen across all positions independently. Such an occurrence is multiplied by the chance both K_1 terms will not be in error, $(1\text{-}e)^2$. The leftmost $(1/p_1)^{n1}$ term represents the possibility of K_1 of two non-matching records equaling when one K_1 can become another only via the error e, which is again the size of the namespace. The e and (1-e) terms signify a mistake occurred in one but not the other field. The factor 2 appears since this can happen for the combination R_i and R_j or the opposite combination event, R_j and R_i. Therefore,

$P\ (R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match}) = ((1/p_1)^{n1})*(2*e*(1\text{-}e) + (1\text{-}e)^2 + e^2) =$
$((1/p_1)^{n1}) * ((1\text{-}e) + e)^2 =$
$((1/p_1)^{n1}) * (1)^2 =$
$(1/p_1)^{n1} \qquad (8)$

Consequently,
$P\ (R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match}) = 1 - (1/p_1)^{n1} \approx 1 \qquad (9)$
That is, the second term is negligible when the namespace p_1^{n1} is assumed large. Thus,
$[K_1]W_{agree} \approx \log(\,(1 - 2*e)\,/\,(1/p_1)^{n1}) = \log((p_1^{n1})*(1 - 2*e)) \qquad (10)$
$[K_1]W_{disagree} \approx \log((2*e)\,/\,1\,) = \log(2*e) \qquad (11)$

We should note that (10) is typically a large positive value in actual record linkage systems since it is a log of a term including a large namespace; (11) is usually a negative, but sometimes a small positive value since it is a log of a term including a small error e. Both are typically notably different from 1 to improve matching.

3.2.2.3.3 Improving Linkage with a Second Field

What criteria must K_2 meet to reduce linkage errors given a system only using a suboptimal K_1? K_2 must have a larger namespace or smaller error rate which will create larger weights to reverse K_1-generated false negatives and false positives. Imagine K_1 is creating a false negative. K_2 must push the total weight above a new threshold. It must fix the mistake when $R_i[K_2] = R_j[K_2]$ but $R_i[K_1] \neq R_j[K_1]$ when R_i and R_j match. When only K_1 was used, such an error would lead to a false negative. Upon encountering $R_i[K_1]$ and $R_j[K_1]$ in a K_1-only system, an error in either K_1 field would set $W_{total\ weight}$ to $[K_1]W_{disagree}$ and the records would be incorrectly labeled.

The threshold for a K_1-only system is

$T1 = ([K_1]W_{agree} + [K_1]W_{disagree}) / 2$

In a K_1-only system we have,
$W_{total\ weight} = [K_1]W_{disagree} < T1 = ([K_1]W_{agree} + [K_1]W_{disagree}) / 2$

The left term would be a low or negative number which would be smaller than the positive value on the right, a positive number plus half of the low or negative term on the left.

For a K_2 system, new weights and threshold must be created since the new field, K_2, is introduced. K_2 parameters are similar to that of a K_1-system:
$[K_2]W_{agree} \approx \log((p_2^{n2})*(1 - 2*f))$ (12)
$[K_2]W_{disagree} \approx \log(2*f)$ (13)

The new threshold is
$T2 = ([K_1]W_{agree} + [K_2]W_{agree} + [K_1]W_{disagree} + [K_2]W_{disagree}) / 2$

To fix K_1 false negatives, we must have
$W_{total\ weight} = [K_1]W_{disagree} + [K_2]W_{agree} \geq T2$ (14)

Thus, K_2 characteristics must meet (14) to fix K_1-created false negatives.

Imagine K_1 is creating false positives. K_2 characteristics must meet additional criteria beyond (14). Upon encountering $R_i[K_1] = R_j[K_1]$, when the records are a non-match, the $W_{total\ weight}$ would be incorrectly set to $[K_1]W_{agree}$ and the records would be incorrectly labeled a match.
$W_{total\ weight} = [K_1]W_{agree} \geq T1 = ([K_1]W_{agree} + [K_1]W_{disagree}) / 2$

The left term would be a positive number, which would be higher than the smaller positive value on the right. In this case, K_2 must create a sufficiently low weight to push the total weight below the threshold. In this case, we must have
$W_{total\ weight} = [K_1]W_{agree} + [K_2]W_{disagree} < T2$ (15)

Combining (14) and (15) we have,
$[K_1]W_{agree} + [K_2]W_{disagree} < T2 \leq [K_1]W_{disagree} + [K_2]W_{agree}$
$\Rightarrow [K_1]W_{agree} + [K_2]W_{disagree} < [K_1]W_{disagree} + [K_2]W_{agree}$ (16)

Reinstating the original parameters, combining, and exponentiating both sides, we have
$\log((p_1^{n1})*(1 - 2*e)) + \log(2*f) < \log(2*e) + \log((p_2^{n2})*(1 - 2*f))$
$\Rightarrow (p_1^{n1})*(1 - 2*e)*(2*f) < (2*e)*(p_2^{n2})*(1 - 2*f)$ (17)

Given an e, p_1, and n1 associated with K_1 we can compute if a given available K_2 can correct for K_1 errors. It has to meet equation (17). We can analyze K_2's parameters when training the system. The best K_2 among available fields can be chosen to fix the platform's linkage errors due to only K_1. Observe that the K_2 error rate, f, can be even higher than the error rate for K_1, e, as long as its namespace compensates for it and allows K_2 to fix K_1's errors.

We can ensure that K_2 is introducing fewer linkage errors than K_1. K_2 is fixing K_1 errors, but ideally it can fix more errors than it introduces in which case overall system linkage would improve. Assume a K_2 field meets (17). We check if

	(*To clarify, in this inequality we are computing:*
$[\,(\,P\,(R_i[K_2] = R_j[K_2] \mid R_i \text{ and } R_j \text{ don't match}) \ast$	[(prob(K_2 false positive) *
$P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})\,) +$	prob(K_1 true negative)) +
$(\,P\,(R_i[K_2] \neq R_j[K_2] \mid R_i \text{ and } R_j \text{ match}) \ast$	(prob(K_2 false negative) *
$P\,(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ match})\,) +$	prob(K_1 true positive)) +
$(\,P\,(R_i[K_2] \neq R_j[K_2] \mid R_i \text{ and } R_j \text{ match}) \ast$	(prob(K_2 false negative) *
$P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match})\,) +$	prob(K_1 false negative)) +
$(\,P\,(R_i[K_2] = R_j\,[K_2] \mid R_i \text{ and } R_j \text{ don't match}) \ast$	(prob(K_2 false positive) *
$P\,(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})\,)\,] <$	prob(K_1 false positive))] <
$P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) +$	prob(K_1 false negative) +
$P\,(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})$	prob(K_1 false positive))

The left part of the inequality represents the full linkage errors produced by a K_1 and K_2 system, while the right part represents the complete errors in a K_1-only system. We hope for the left side to be smaller than the right to create fewer errors. We simplify the above inequality:

$[\,P\,(R_i[K_2] = R_j[K_2] \mid R_i \text{ and } R_j \text{ don't match}) \ast$	[prob(K_2 false positive) *
$(P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match}) +$	(prob(K_1 true negative) +
$P\,(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})\,) +$	prob(K_1 false positive)) +
$P\,(R_i[K_2] \neq R_j[K_2] \mid R_i \text{ and } R_j \text{ match}) \ast$	prob(K_2 false negative) *
$(P\,(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) +$	(prob(K_1 true positive) +
$P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match})\,)\,] <$	prob(K_1 false negative)) <
$P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) +$	prob(K_1 false negative) +
$P\,(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})$	prob(K_1 false positive)
➔ $[P(R_i[K_2] = R_j[K_2] \mid R_i \text{ and } R_j \text{ don't match})\ast(1) +$	[prob(K_2 false positive)*(1)+
$P(R_i[K_2] \neq R_j[K_2] \mid R_i \text{ and } R_j \text{ match})\ast(1)\,] <$	prob(K_2 false negative)*(1)] <
$P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) +$	prob(K_1 false negative) +
$P\,(R_i[K_1] = R_j[K_1] \mid R_i \text{ and } R_j \text{ don't match})$	prob(K_1 false positive)
➔ $[\,P\,(R_i[K_2] = R_j[K_2] \mid R_i \text{ and } R_j \text{ don't match}) +$	[prob(K_2 false positive) +
$P\,(R_i[K_2] \neq R_j[K_2] \mid R_i \text{ and } R_j \text{ match})\,] <$	prob(K_2 false negative)] <
$P\,(R_i[K_1] \neq R_j[K_1] \mid R_i \text{ and } R_j \text{ match}) +$	prob(K_1 false negative) +

$$P\ (R_i[K_1] = R_j[K_1]\ |\ R_i \text{ and } R_j \text{ don't match}) \qquad \textit{prob}(K_1 \textit{ false positive}) \quad (18)$$

We insert the approximations from (6) and (8) for K_1 above, and analogous ones for K_2, into (18):

$$(1/p_2)^{n2} + 2*f < 2*e + (1/p_1)^{n1} \qquad (19)$$

The total linkage error rate in a K_1-only system is the sum on the right. The 2*e term is the error rate when K_1 is not equal during a match while the $(1/p_1)^{n1}$ term is the error rate when K_1 is equal during a non-match. The sum on the left is the equivalent error rate for a K_2- and K_1-system, using analogous computations as for K_1. K_2 characteristics are apparently the only ones responsible for linkage errors. The values for errors are analogous to those computed for a K_1-only system. The K_2 field among all candidate K_2 fields with the smallest product on the left in the above inequality should be chosen to maximally overcome identifier errors in the system.

3.2.2.3.4 Assessing a Field's Characters for Improved Linkage

We examine the information value of another common pattern used in record linkage: the analysis of a field's characters. Can linkage problems due to identifier errors be reduced if a field's characters are examined? If the number of character positions in error in a field is not large, the field's characters may be analyzed. More refined weights will be returned which can fix some linkage errors. However, if a field is diverse and dense, more false positives may be created due to "close" fields. A greater number of higher weights may be created which would be larger than the threshold than the errors fixed by the character level analysis, overcoming the benefits of such analysis. We will work with string comparators as the mechanism for character-level assessment. We will also continue to work with the agreement and disagreement weights due to the frequent usage of equality and inequality patterns in practice. A weight can be created by a string comparator just like the agreement and disagreement weights.[439] Since a string comparator ultimately indicates degree of agreement between fields, the weight attributed should fall between $W_{disagree}$ and W_{agree} to reflect the comparator's score, reflecting the degree of field equality. One author converted the comparator score into [0,1] for simpler data manipulation and analysis, which we also do for similar reasons.[440] The value of zero should indicate virtually no similarity between fields while a 1 can indicate an exact match between fields. A score close to 0 should be transformed into a weight close to $W_{disagree}$, reflecting almost full field inequality. A converted score close to 1 should be transformed into a weight close to W_{agree}, reflecting almost full field equality. A simple model to create the needed comparator weight is to assume that the rise from the disagreement to agreement weights is linear. The [0,1] score can simply be multiplied by the difference between the disagreement and agreement weights and added to $W_{disagree}$.

[439] Taken from William Yancey, "An Adaptive String Comparator for Record Linkage," 19 February 2004, 1-24, <http://www.census.gov/srd/papers/pdf/rrs2004-02.pdf> (24 May 2005).

[440] For instance, see Grannis, "Real World Performance of Approximate String Comparators for Use in Patient Matching," 44.

A more advanced weight construction would involve examining the training file, investigating particular [0,1] scores. Linear regression can be used to create a function mapping the converted scores into appropriate weights representing the comparator score. The analyst can find how often a converted comparator score in [0,1], when examining field K among file records, equaled some particular score z in [0,1] when the records matched as opposed to how often such scores equaled score z when the records didn't match. The values would be quantified for all training file scores z. We elaborate on the analysis. Imagine that a string comparator *compare(s1, s2)* returns a degree of equality score between strings s1 and s2. If *compare(s1, s2)* can be converted into $0 \leq z \leq 1$, the weight for comparing arbitrary field K in records R_i and R_j in our context becomes:

$$W_{comparator} = \log\left(P\left(\text{convert}\left(compare(R_i[K], R_j[K])\right) = z \mid R_i \text{ and } R_j \text{ match}\right) / P\left(\text{convert}\left(compare(R_i[K], R_j[K])\right) = z \mid R_i \text{ and } R_j \text{ don't match}\right)\right)$$

$W_{comparator}$ should compute to a high positive value when there are more matches for a given score z in [0,1] than non-matches. Likewise, $W_{comparator}$ should compute to a low or negative value when there are more non-matches for a given score z than matches. Linear regression can be used to fit a line between $W_{disagree}$ and W_{agree} representing the points (z_i, $W_{comparator}$ [z_i]) for all scores z_i found and examined in the training file. The resulting function $W_{comparator}$ would compute the weights for all scores z when examining the validation file. The system would obtain a regular comparator score when examining $R_i[K]$ and $R_j[K]$ in a validation file, convert it to [0,1], and convert this value into the needed weight using the $W_{comparator}$ function. All other described record linkage processes, computing $W_{total\ weight}$, comparing it to a threshold, etc., are the same. We will use the former linear-based weight-creation approach above for simpler presentation. However, our analysis will work with more advanced weight constructions.

We show how incorporating the Levenshtein string comparator into one field at first appears useful to reduce linkage identifier errors. Our results will be demonstrated numerically. We will work with a system using K_1 and K_2 as before. The parameters for our hypothetical system are as follows:

$p_2 = p_1 = 10$
$f = e = 0.01$
$n1 = 8$
$n2 = 12$

The characteristics of K_1 and K_2 are similar. As is evident, the length of K_2, n2, is 4 positions longer than that of K_1, n1.

Incorporating a comparator can correct the false negatives created by the same field. With the above choice of parameters, K_2 can correct K_1 errors. K_2 meets equation (17).
$(p_1^{n1})*(1 - 2*e)*(2*f) < (2*e)*(p_2^{n2})*(1 - 2*f)$
➔ $(10^8)*(1 - 2*(0.01))*(2*(0.01)) < (2*(0.01))*(10^{12})*(1 - 2*(0.01))$
➔ $1.96*10E+6 < 1.96*10E+10$

Further, K_2 should be introducing fewer overall system linkage errors as it meets equation (19).
$(1/p_2)^{n2} + 2*f < 2*e + (1/p_1)^{n1}$
➔ $(1/10)^{12} + 2*(0.01) < 2*(0.01) + (1/10)^{8}$
➔ $0.02 + 1.0*10E\text{-}12 < 0.02 + 1.0*10E\text{-}8$

Since both inequalities hold K_2 is a useful field for the system.

Imagine that the errors produced by f create a single typo in K_2. That is, the K_2 error is a single spelling mistake in one character. The comparator is useful in K_2 to fix K_2 false negatives. Suppose that R_i and R_j match. If LEV represents the Levenshtein string comparator function call, we have
$\text{LEV }(R_i[K_2], R_j[K_2']) = 1$

K_2' is a K_2 field which has been erroneously transformed. The *distance* between K_2' and K_2 is one character position, which has been changed due to the spelling mistake. We first examine the case when both K_2 and K_1 are in error. In this case, we hope the comparator can reverse the false negatives if the K_2 error is small. The weight produced should be high to reverse the K_2- and K_1-generated false negatives. The weight assigned to K_2 should push the total weight above the threshold. We have,
$T2 = ([K_1]W_{agree} + [K_2]W_{agree} + [K_1]W_{disagree} + [K_2]W_{disagree}) / 2$

The Levenshtein string comparator score can be transformed into a modified [0,1] score following Grannis:[441]
$z_{score} = 1 - [\text{LEV}(s1, s2) / \text{maxlen}(s1, s2)]$

We have,

$$\begin{aligned} z_{score} &= 1 - [\text{LEV}(R_i[K_2],R_j[K_2']) / \text{maxlen }(R_i[K_2],R_j[K_2'])] \\ &= 1 - (1 / n2) \\ &= 1 - (1 / 12) \\ &= 0.92 \end{aligned}$$

The comparator weight is created from the linear difference between the agreement and disagreement weights:
$[K_2]W_{comparator} = [K_2]W_{disagree} + ([K_2]W_{agree} - [K_2]W_{disagree}) * z_{score}$

The total weight is
$W_{total\ weight} = [K_2]W_{comparator} + [K_1]W_{disagree}$

We check if $W_{total\ weight} \geq T2$.

$$\begin{aligned} &[K_2]W_{comparator} + [K_1]W_{disagree} \geq T2 = \\ &\qquad ([K_1]W_{agree} + [K_2]W_{agree} + [K_1]W_{disagree} + [K_2]W_{disagree}) / 2 \end{aligned}$$

[441] Grannis, "Real World Performance of Approximate String Comparators for Use in Patient Matching," 44.

➔ $[K_2]W_{disagree} + ([K_2]W_{agree} - [K_2]W_{disagree}) * z_{score} + [K_1]W_{disagree} \geq$
$([K_1]W_{agree} + [K_2]W_{agree} + [K_1]W_{disagree} + [K_2]W_{disagree}) / 2$
➔ $\log(2*(0.01)) +$
$(\log((10^{12})*(1 - 2*(0.01))) - \log(2*(0.01))) * (0.92) + \log(2*(0.01)) \geq$
$(\log((10^{8})*(1 - 2*(0.01))) + \log((10^{12})*(1 - 2*0.01)) +$
$\log(2*(0.01)) + \log(2*(0.01))) / 2$
➔ $(-5.64) + (39.83 - (-5.64)) * (0.92) + (-5.64) \geq (26.55 + 39.83 + (-5.64) + (-5.64)) / 2$
➔ $30.55 \geq 27.55$

Since the total weight rises above the threshold, K_2 can correct its own false negatives when K_1 is in error.

K_2 will also correct its own false negatives when K_1 is error-free. $[K_1]W_{agree}$ will be added to $[K_2]W_{comparator}$ which will make the sum above on the left even larger than 30.55, and thus larger than the threshold, 27.55, generating the match. The Levenshtein string comparator appears useful.

However, the comparator will also be assigning the same sum 30.55 to many other "close" non-matching fields in dense space. When K_2 is dense and diverse, as was one of our earlier assumptions, for a given field there should be a number of fields which may differ by one character. There may be many non-matching identifiers which would differ in any one of K_2 n2 positions. The weight for such identifiers will be the same, 30.55, in the case when the K_1 fields and K_2 fields are not equal. K_2's small error rate f will probably generate considerably fewer false negatives as it is probably smaller than the number of "close" fields in K_2's namespace. In this example, performing character-level analysis for K_2 will not be useful and should not be incorporated as the assigned total weight will be larger than the threshold. More system linkage errors will be introduced than fixed.

The same analysis above can be carried out to determine which error rate, length of field, and number of characters per position is needed for character-level analysis to be useful for a field. Future research can create the needed approach.

3.3 Threat Framework

We ask how to handle the above record linkage paradigm securely. We return to the other part of Enhancement 1 and ask how to deidentify the above operations. This thesis proposes to securely link records, including allowing for character-level analysis. To do so we create a security framework. A variety of security approaches exist from which we must select the best foundation. We synthesized a literature analysis with interviews with experts to derive the following *security and functionality framework*. Approaches meeting the *security and functionality framework* may be used for securing the above linkage operations; those not meeting it may be rejected.

- Efficiency is obviously preferred. If two technical approaches offer the same level of functionality and security, the one which uses less storage, less time, etc., is obviously preferred.
- Hiding smaller "secrets" is preferred. Hiding a very small amount of data, for instance, one bit, can be considered insecure since the data can be guessed despite the strongest of security. Nevertheless, given a certain nominal amount of data which are difficult to guess, protecting less of such data is easier than protecting more of them as less opportunity exists for misuse. For example, some authors critique access controls as compared to encryption from this perspective within a database context.[442] [443] Data are ultimately unencrypted when access controls in a database are used, creating more difficulty in controlling plaintext management and disclosure. An encryption approach restricts plaintext availability more as the secret is placed in a small encryption key which alone should be protected.
- If an approach involves encryption, the encryption algorithm itself—the mathematical formula(s) and any public parameters—should be public knowledge.[444]
- Any encryption-based approach must be secure. As we will investigate encryption and hashing in this thesis, we will create a threat model for investigating the associated cryptographic schemes in an intra- and inter-organizational context. The BQMA are internal applications; however, record linkage can happen outside of an organization, too, if the organization sends its data elsewhere for analysis. The approach must be secure in both contexts. The threat model will be described in more detail in Section 3.3.1 below.
- It should not be feasible to conduct brute force attacks (BFAs) on the approach either. In this thesis, BFAs will be defined as systematic traversals of all possibilities until "success" is obtained.[445] We will define the nature of BFAs in Section 3.3.2.

3.3.1 Cryptographic Threat Model

We first describe the cryptographic threat model against which our approach should be secure because we will explore a cryptographic solution in this thesis. Over the years,

[442] Min-Shiang Hwang and Wei-Pang Yang, "A Two-phase Encryption Scheme for Enhancing Database Security," *Journal of Systems Software*, 31 (1995): 257-265.

[443] George I. Davida and David Wells, "A Database Encryption System with Subkeys," *ACM Transactions on Database Systems*, 6 (1981): 312-328.

[444] Indeed, in the cryptographic world it is a typical assumption that any algorithm to be examined is public. Fewer faults might be uncovered in proprietary approaches due to less scrutiny by experts. (See Bruce Schneier, "Security Pitfalls in Cryptographic Design," *Information Management & Computer Security*, 6 (1998): 133-137; also, see RSA Security, "What is Cryptanalysis?" *RSA Laboratories' Frequently Asked Questions about Today's Cryptography*, <http://www.rsasecurity.com/rsalabs/node.asp?id=2200> (14 March 2004)).

[445] Extracted from Terry Ritter, "Brute Force Attack," *Ritter's Crypto Glossary and Dictionary of Technical Cryptography*, 12 March 2004, <http://www.ciphersbyritter.com/GLOSSARY.HTM#BruteForceAttack> (15 March 2004).

cryptographers have designed threat models to capture ways in which an adversary can interact with an encryption or hashing system to capture sensitive data. Consider one common high-level general classification of attacks on encryption or hashing schemes, in order, approximately, of least to most severe as presented by Menezes.[446] [447] [448] In the attacks below the purpose of the adversary is to deduce any involved encryption or hashing algorithm's key, or surmise the current, past, or future plaintexts associated with the system. If an adversary is successful, we call such success system *compromise.*

1) A *ciphertext-only attack* is an attack in which the adversary tries to compromise the system by observing only available ciphertexts. He may try to learn the distribution of ciphertexts or the nature of the encryption or hashing encodings in hopes of learning more about the original plaintexts.
2) A *known-plaintext attack* is one in which the adversary has a quantity of plaintext and corresponding ciphertext. He may uncover more sensitive data or attempt other compromises knowing such mappings.
3) In a *chosen-plaintext attack,* the adversary chooses some amount of plaintext and is given the corresponding ciphertext. He is forcing the system to encrypt or hash data, hoping to discover the nature of the encryption or hashing process, which may be sensitive to such mapping.
4) An *adaptive chosen-plaintext attack* is a *chosen-plaintext attack* wherein the adversary may choose the plaintext to be given based on the ciphertexts received from prior requests. He examines chosen plaintext and ciphertext structures dynamically, "testing" the system with probing requests.
5) A *chosen-ciphertext attack* is one in which the adversary selects the ciphertexts and is given the corresponding plaintexts. He begins with the ciphertexts, generates the plaintexts, and again tries to surmise the encryption or hashing transformations. Practically, such an attack may happen if the adversary gains access to the decryption equipment; however, the assumption is he will not access the decryption key.[449]

[446] Menezes, "Chapter 1," 41.

[447] See Menezes, "Chapter 1," 42.

[448] Alfred Menezes, Paul C. van Oorschot, and Scott A. Vanstone, "Chapter 9," *Handbook of Applied Cryptography*, 2001, 326, <http://www.cacr.math.uwaterloo.ca/hac/about/chap9.pdf> (24 May 2005).

[449] If the adversary can get access to the key, the system is compromised and the application owner has to "recover." For example, he may need to warn data owners that their data has been compromised, etc. Some cryptographic designs try to protect the key. (For example, see Freesoft.org, "Connected: An Internet Encyclopedia: Key Management," <http://www.freesoft.org/CIE/Topics/138.htm> (24 May 2005)). For example, they can use passwords to protect the key or split the key to make the construction of the key more difficult. If such approaches work in our record linkage paradigm they can be chosen for improved key security. However, if such approaches cannot be used in our context due to, for example, architectural incompatibility, we assume that approaches that protect the key less will be chosen. Standardized robust encryption schemes protect the key, such as recommending the key not be given to colleagues or that the key be stored in a secure location. Data subject to such schemes will be protected, albeit less securely than if extra key protection could be involved. Also, with respect to hash functions, decryption can't be done per se. The purpose of hashing is to confirm data integrity, not to recover original data. (See discussion in Menezes, "Chapter 9," 321-322). The decryption attacks described in the text would not be possible; only plaintext-based attacks could be mounted. Enhancement 1 requires a "decryption" for data subject reidentification to be available. We will examine one approach which tries to deidentify record linkage applications later in this thesis to understand how hashing may "decrypt."

6) An *adaptive chosen-ciphertext attack* is a *chosen-ciphertext attack* wherein the choice of ciphertexts may depend on the plaintexts received from prior requests. Such an attack starts with "testing" ciphertexts and obtaining, hopefully, useful plaintexts to understand the encryption or hashing transformations.

Each of these attacks can be formalized in theoretical terms.[450] For example, to test the security of an encryption or hashing scheme E, an "oracle," i.e., a program, is given access to and can operate E, represented as another program. The adversary, a third program, attempts to compromise E by interacting with the oracle. The security of E can be expressed as the adversary's "advantage," the time and space complexity of the oracle and the adversary involved, in being able to divulge the key, discover some plaintexts, etc. as described in the *compromise* definition above.

We will not be using the theoretical versions of such attacks but the practical realization of such attacks within and outside of institutions. More pragmatic threat models for actual software operations will be constructed. With respect to the intra-organizational context, i.e. internal attacks, for example, we will be guided by the "insider threat" discussion from Section 1.3.2. In that discussion, some employees will have access to some of the equipment, personnel, or even some of the data used in the encryption or hashing processes some of the time. They may encrypt or decrypt or hash data to obtain needed information, trying to misuse a cryptographic system. In a real-world *known-plaintext attack*, for instance, an employee might know how original sensitive data becomes particular encrypted output. The employee might be a business manager who generated such data. He subsequently gives it to an IT staff specialist for secure storage. Perhaps the data will be examined later on. He sees the ciphertexts that are created and stored in a data warehouse because the IT expert protects them with encryption, storing them in such a warehouse. Since the original employee is a business manager he might have access privileges to the data warehouse for operational reasons. To the degree the encryption used by the IT staff specialist is the same as that used by the record linkage application we're exploring (perhaps the encryption process is standardized within the organization), the business manager has just conducted a realistic "known-plaintext attack." He knows the relationships between some plaintexts and corresponding ciphertexts because he sees the latter in the data warehouse. Other similar attacks within and outside of institutions may happen, too.

We call our threat framework *attacks against encryption* (AAE). This framework will encompass the six attacks described above, practically realized. The adversary will be an employee running the record linkage software, attempting system compromise. The plaintexts will be the application fields, such as a policy-holder's last name or Social Security Number which could be available in a linkage data set. The ciphertexts will be these encrypted or hashed fields available to the record linkage software and to the employee running it.

[450] This is taken from Shafi Goldwasser and Mihir Bellare, *Lecture Notes on Cryptography* (Cambridge, MA: Massachusetts Institute of Technology, August 2001), 62, 91-93.

We need to clarify our notion of compromise. In addition to learning past, current, or future plaintexts, or even the key involved with an encryption or hashing system, the discovery of small parts of a plaintext will also be considered a compromise under the AAE. The full plaintext might be guessed if such partial knowledge becomes known. As an example of practical guessing, imagine the linkage data set is like the copy data store, a claims data set, deidentified via Safe Harbor. An organizational employee is trying to guess the personal information of policy-holders using their *last names*. If a policy-holder associated with a claims record has a rare last name, knowing its first few characters can considerably help in guessing that person's fully identifying information. A phone book can be used covering the zip code of the policy-holder, which will be available in the claims record but will now be three digits long after Safe Harbor deidentification. Several last names might be listed in the phone book matching the first few characters of the policy-holder's last name. Inputting the several last names and the patient's state, represented by the three-digit zip code, into a free online name lookup service such as www.MelissaDATA.com will yield the first name, last name, and current age of all individuals with that last name within available public and other records in that state available to the lookup service.[451] Age can be computed from the claims data by subtracting the year of birth field, available from the Safe Harbor-deidentified date-of-birth, from the current year. Matching the age with the age provided by the lookup service can narrow the search to the possibly correct person. I was able to successfully obtain, within 30 minutes, names, home addresses, and phone numbers for a few people with rare last names in this manner knowing only the first few last name characters based on various lookup strategies using www.MelissaDATA.com and similar free online lookup services.[452] [453]

[451] www.MelissaDATA.com, "People Finder Lookup," <http://www.melissadata.com/cgi-bin/peoplefinder.asp> (20 October 2005).

[452] Lookups performed on October 20, 2005.

[453] As another example of being able to guess personal information, consider the Social Security Number (SSN). In the PM context, the medical record number on a claims record may be an SSN. The Federal Tax Number can be an SSN on a UB92, for instance. (Centers for Medicare and Medicaid Services, UB92; Centers for Medicare and Medicaid Services, HCFA 1500). However, numbers making up the SSN are not randomly assigned. The first three SSN digits comprise the "area" of the SSN. This value is determined by the zip code of the mailing address of the individual who submitted the application for the SSN. (Social Security Administration, "Is There Any Significance to the Numbers Assigned in the Social Security Number?" 2003, <http://ssa-custhelp.ssa.gov/cgi-bin/ssa.cfg/php/enduser/std_adp.php?p_sid=mtTFC74h&p_lva=&p_faqid=87&p_created=955483216&p_sp=cF9zcmNoPSZwX2dyaWRzb3J0PSZwX3Jvd19jbnQ9NjImcF9jYXRfbHZsMT0xNiZwX3BhZ2U9MQ**&p_li> (14 March 2004)). The Social Security Administration (SSA) website indicates that digit 5 is the most popular first digit in the "area" field. (Social Security Administration, "Social Security Number Allocations," <http://www.socialsecurity.gov/foia/stateweb.html> (14 March 2004)). Imagine that a deterministic encryption function is used to encrypt such values. Deterministic encryption transforms the same plaintexts into the same ciphertexts every time, as will be described later on in the text. If the first character position of the plaintext is enciphered on its own, the resulting ciphertext of the first position character will have the same frequency distribution as digit 5 across all enciphered SSNs. Deterministic encryption does not change the frequency of the underlying data. The created ciphertext code will have the same frequency, allowing for the recognition of an original plaintext character based on the frequency distribution of the corresponding ciphertext code. The SSA describes other smaller "structures" that may exist within the SSN, which could permit other types of attacks based on smaller partial plaintext knowledge, which, in turn, may allow for better full-field SSN guesses.

3.3.2 Brute Force Attacks

Next we describe the brute force attacks (BFAs) against which an approach must also protect. Just as in the AAE above, success via a BFA will involve the discovery of sensitive data. If cryptography is involved, the discovery of full or partial original plaintexts or the encryption or hashing key in key-based schemes would be a compromise.[454] [455] We divide the BFAs into a *plaintext-brute force attack* and *key-brute force attack*, against both of which an approach must guard. To say the BFAs should be "infeasible," as described before, will mean the storage or time requirements for attacking the scheme would make the BFAs unrealistic given today's available computational resources. The design of good encryption or hashing schemes is to make the BFAs some of the few possible attacks on a scheme yet make the attacks virtually impossible to carry out.

We describe the *plaintext-brute force attack* and *key-brute force attack*. In a *plaintext-brute force attack*, the adversary investigates every possible plaintext in an attempt to compromise the scheme. He traverses all possible plaintexts in the domain and matches the results to those which have been encrypted or hashed and to which he has access through normal system interactions. For example, in a PM context, assume medical record numbers (MRN) are to be encrypted. Suppose MRN values are arbitrarily in the range of 0000000000 – 9999999999. We define a *plaintext-brute force attack* as the success an employee has by systematically encrypting 0000000000 – 9999999999 to see if some enciphered member id equals one of the enciphered member ids in the copy data store to which he has access. If this can be done for all or some of the full original MRNs, or all or some of partial MRNs (i.e., particular character positions), the scheme is insecure. We define the *key-brute force attack* in a similar fashion. The adversary systematically generates all possible keys and encrypts or hashes the known plaintext field values to see if the ciphertexts in the copy data store to which he has access result. The adversary, based on the AAE, should already have some such plaintext and ciphertext pairs, such as from a *known-plaintext attack* in an intra-organizational context. He can use such data to systematically traverse through all possible keys of a key-based scheme to find the right key.

Whether the adversary can mount such an attack actually depends on the design of the scheme. In some constructions, it would be impossible to find the right key. An employee should only have a limited set of possible plaintexts and corresponding ciphertexts per the AAE.[456] However, a construction may allow for many keys to map the available plaintexts into their corresponding ciphertexts, undermining the recognition of the right

[454] For example, Menezes, "Chapter 1," 42.

[455] Menezes, "Chapter 9," 336.

[456] Otherwise there would be broader security concerns within the organization. Large amounts of sensitive data are available to unauthorized individuals.

key.[457] Some encryption approaches create a key space that includes all possible transformations from the domain of the key space into its range. For example, a simple substitution cipher could have the same domain and range. The domain and range can all be a list of items, each of which is valued, say arbitrarily, 0 to N. The cipher works by permuting a given domain value into a different value in the range.[458] In such a construction the key space is at least N! in size. Each domain value maps into any range value except those that were already mapped. A *key-brute force attack* could not identify the correct key. The at least N! set of permutations representing the keys by their very nature would create many possible "keys" that could map the small set of known plaintexts into the corresponding ciphertexts. The proper key cannot be identified since many keys would transform available plaintexts into corresponding ciphertexts for that plaintext-ciphertext space since their transformations over that space may be identical.

However, a *key-brute force attack* can be mounted on approaches which dramatically limit the key space. For example, AES, a block cipher, can be the basis of deterministic encryption schemes.[459] AES has a much smaller key space than the substitution cipher described above. AES breaks all input into 128-bit plaintext blocks before encrypting. The typical ciphertext blocks produced are each 128 bits in size.[460] The key space used for AES is 2^{128} or at most 2^{256} bits in size, it is not $2^{128}!$ in size.[461] If the wrong key is

[457] Note, hash functions typically convert a larger domain into a smaller range. Therefore, for keyed hash functions, more than one key might be found which maps a given element of a domain into the same element in the range. For example, if only one plaintext-ciphertext pair is known via a *known-plaintext attack*, many keys can be found. (See Menezes, "Chapter 9," 336). They would map the single plaintext into the single available ciphertext. The way to find the right key in a *key-brute force attack* on a keyed hashing scheme is to have enough plaintext-ciphertext pairs via a *known-plaintext attack* or similar attacks to disqualify all keys but one. Given the nature of the insider threat, we'll assume there are enough such pairs available. If a hashing approach implements a key structure requiring innumerable known-plaintext pairs to disqualify keys, the hashing algorithm will be considered safe from a *key-brute force attack*. Although some plaintext-ciphertext pairs will be known to an employee via the AAE, the number of pairs available should not be as large as needed for attack success.

[458] See Menezes, "Chapter 1," 15-17.

[459] See Goldwasser, *Lecture Notes on Cryptography*, 51-53, 84.

[460] National Institute of Standards and Technology, "Announcing the Advanced Encryption Standard," 26 November 2001, 7, <http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf> (18 October 2005).

[461] Indeed, such constructions are often based on pseudo-random functions (in case of AES, it is based on the similar pseudo-random *permutation*), which are a family of functions such that a randomly-chosen function is "computationally indistinguishable" from a truly random function. (Goldwasser, *Lecture Notes on Cryptography*, 67, 61). Such constructions must restrict their key space to make the algorithm behave like a pseudo-random function. Computational indistinguishability implies that a given plaintext should typically map into a given ciphertext only under one or a small set of keys. Such functions should typically send the same plaintexts into random ciphertexts under different keys, i.e., "random" behavior. A randomly selected key for the algorithm is part of the randomness making the deterministic algorithm like a pseudo-random function. (See Yehuda Lindell, "Pseudorandomness and Private-key Encryption Schemes," *Introduction to Cryptography*, <www.cs.biu.ac.il/~lindell/89-656/lecture03-89-656.ps> (31 August 2005); Wikipedia, "Key (Cryptography)," 14 July 2005, <http://en.wikipedia.org/wiki/Key_(cryptography)> (31 August 2005)). If the key space is not reduced and the wrong key is used, the function may continually map known plaintexts into available ciphertexts. Many keys may be available for such mappings. As there are numerous keys, some would only transform the domain space excluding the plaintexts available via the AAE. For the plaintexts discovered via the AAE, the keys would create the same plaintext-to-ciphertext transformations, as in that space they can create identical mappings, as described in the text. This would

used in such constructions, this should be recognized in a *key-brute force attack.*[462] The key would map known plaintexts into non-corresponding ciphertexts. If the right key were found it would map known plaintexts into corresponding ciphertexts. All unique original plaintexts could be obtained since by definition, decryption associated with any encryption scheme must transform ciphertexts into the original plaintexts. If an approach is implemented in this manner permitting such an attack it is obviously insecure.

3.4 Securing Record Linkage

Given the above *security and functionality framework*, we build a secure approach that meets it. We devise a key-based encryption approach to securely address errors in linkage identifiers, as per Enhancement 1. We create a general solution that can incorporate many fields and allow for character-level analysis of a field. First, we describe how to secure a full field. To secure a full field deterministically encrypt it. Equality of ciphertexts can be directly compared since the same ciphertexts should be created for the same plaintexts under deterministic encryption which creates the same ciphertexts from identical plaintexts.[463] With regard to character-level analysis, we will use the following example for illustration. Assume that originally the system is using just one field for linkage and needs to analyze its characters. For example, perhaps if all positions in the identifier equal but a particular position does not, this signifies the individual is a dependent of the policy-holder. The intent is to link all family-related information. Another field in the records will be used for *padding* as will be described below. We construct a scheme involving both fields, permitting character-level analysis of the first field. First, choose a deterministic key-based encryption algorithm which can withstand *attacks against encryption* (AAE). This same approach should also withstand a *key-brute force attack.* Call this the *foundation algorithm.* AES, which can be the basis of deterministic schemes, can be a block cipher meeting these criteria, although subject to the following important caveat.[464] Since we will need to equate ciphertexts, as will be shown, if we use block

undermine the function's *pseudo-randomness* because the function is transforming the same plaintexts into the same, not different, ciphertexts, undermining "random" behavior.

[462] Just as in the hash function discussion referenced earlier, in a *key-brute force attack* a small number of plaintext-ciphertext pairs, but enough need to be available from the *known-plaintext* or more advanced AAE attacks to discard essentially all but the right key. (Alfred Menezes, "Chapter 7," *Handbook of Applied Cryptography*, 2001, 233, <http://www.cacr.math.uwaterloo.ca/hac/about/chap7.pdf> (24 May 2005); Alfred Menezes, "Chapter 8," *Handbook of Applied Cryptography*, 2001, 286, <http://www.cacr.math.uwaterloo.ca/hac/about/chap8.pdf> (24 May 2005)).

[463] Wikipedia, "Deterministic Encryption," 20 July 2005, <http://en.wikipedia.org/wiki/Deterministic_encryption> (15 September 2005).

[464] See Goldwasser, *Lecture Notes on Cryptography*, 51-53, 84. AES is secure against the AAE when it is run in the Counter or the Cipher Block Chaining (CBC) operating modes which create randomized ciphertexts for every plaintext, breaking any possible relationship between any two encrypted plaintexts. (See Mike Touloumtzis, "Re: (AES) Loopback Crypto Questions," 11 July 2001, <http://mail.nl.linux.org/linux-crypto/2001-07/msg00150.html> (31 August 2005); Steve Weis, PhD student, Lab for Computer Science, MIT, personal interview, September 9, 2004; Goldwasser, Lecture N*otes on Cryptography*, 84-6, 100-110). AES also protects against a *key-brute force attack.* One analysis

ciphers, which can be run in several modes we will run them in Electronic Code Book (ECB) mode.[465] Under such usage, the scheme transforms the same plaintexts into the same ciphertexts, allowing for easy equality comparisons. Under ECB, no encryption scheme including one based on AES can be considered completely secure against the AAE due to the scheme's deterministic operation. The same plaintexts would create the same ciphertexts, allowing an adversary to learn more "information" about equal plaintexts, that is, in this case, they equal. We will explore the degree of insecurity associated with our use of ECB mode and deterministic encryption more generally later in Section 3.4.1.1.

Assume we are working with fields K_1 and K_2 where K_1 needs to be character-analyzed. To securely enable character-level analysis encrypt K_1 in all records R before they reach the validation file. K_1 will be transformed into a new privacy-preserving data structure which will be stored in the original K_1 location.[466] Break K_1 into its individual characters. If R is a record about to enter the validation file, parse its K_1 into $R[K_1][1]$, $R[K_1][2],\ldots,R[K_1][n1]$, where n1 is the number of character positions in K_1. Concatenate to each $R[K_1][q]$ all of K_2, for q=1…n1. Encrypt each of these plaintexts with the foundation algorithm. Link these ciphertexts in a list. Encrypt K_1 with the foundation algorithm. We assemble our privacy-protecting data structure. The data structure will contain the encryption of the full K_1 identifier and the pointer to the linked list of the identifier's characters padded with K_2. We store the fully encrypted K_1 and the head pointer of the linked list in the data structure. Encrypt the full K_2 plaintext and store the ciphertext in the new K_2 privacy-protecting data structure just as for K_1. Below is an illustration of the new privacy-protecting data structures for arbitrary record R using foundation algorithm E:

Before Privacy Enhancement

Rec.	K_1	K_2
R	578	HL4

After Privacy Enhancement

Rec.	K_1	K_2
R	id: E(578) hd: E(5|HL4)->E(7|HL4)->E(8|HL4)	id: E(HL4)

Each data structure has component "id" corresponding to the fully encrypted identifier, and the character-level analyzed field(s) has a linked list head ("hd") as above.

shows that mounting a *key-brute force attack* on AES using its 128-bit key configuration and possibly testing 255 keys per second would take 149 trillion years for the attack to succeed! (Jim Reavis, Feature: Goodbye DES, Hello AES," *Networkworld*, July 30, 2001, <http://www.networkworld.com/research/2001/0730feat2.html> (24 May 2005)). Note, when we refer to AES in this thesis a 128-bit key will be assumed, one of the possible AES key sizes.

[465] Goldwasser, *Lecture Notes on Cryptography*, 84.

[466] This is obviously implementation-specific and depends on how the system currently stores and manages data. For example, in a database scheme, new database columns may be used to hold the new data structure(s).

We explain how a record linkage application can work with the new field formats. All records in the validation file have been encrypted as above. Imagine that, in one particular matching (R_i,R_j) pair in the validation file, all positions between $R_i[K_1]$ and $R_j[K_1]$ are the same except for one, signifying the special dependent status defined above. Also, $R_i[K_2] = R_j[K_2]$. Seeing the privacy-preserving data structures for K_1 and K_2 in the R_i and R_j validation pair in the validation file, the application begins with K_1. The software compares the identifiers. Since $R_i[K_1.id] \neq R_j[K_1.id]$ the two plaintext fields must differ. The application can see that $R_i[K_2.id] = R_j[K_2.id]$. Therefore, it can proceed with trying to determine if the two fields except for one position match, or if this is a false positive with regard to K_2. The only way the encrypted K_1's would not equal while the encrypted K_2's would equal would be for the original R_i and R_j records to not match, and for a mistake to make the original K_2 plaintexts of both records equal; or for the R_i and R_j records to match and we would need to surmise if one character position in the original K_1 plaintexts differ, which would signify another family member, as before. Of course, if $R_i[K_2]$ or $R_j[K_2]$ is also in error, that is, there is a mistake in one or both K_2 fields, no further processing can take place. The pads are in error as the K_2 fields are erroneous. Comparisons of the enciphered K_1 characters will fail. The application accesses the linked lists connected with $R_i[K_1]$ and $R_j[K_1]$. It confirms that each individual character matches except one and properly designates the two records as linking. The below illustrates such a comparison for records R_i and R_j using the example above where K_1=578 and K_2=HL4. Although the ***last digit*** of $R_i[K_1]$ is different, the algorithm can still link R_i and R_j:

Rec.	K_1	K_2
R_i	id: E(57***2***) hd: E(5|HL4)->E(7|HL4)->***E(2|HL4)***	id: E(HL4)
R_j	id: E(578) hd: E(5|HL4)->E(7|HL4)->E(8|HL4)	id: E(HL4)

Other variables, including their character-by-character comparison, can be used, too. The theoretical notions of the information value of fields we described through all of Section 3.2.2 can be used as well. Both security and matching should improve.

Reidentification of units in the validation file is possible via the key from the foundation algorithm. The key is used to decrypt the full linkage fields from the matched records, e.g., the $R[K_1.id]$ or $R[K_2.id]$, as needed.

3.4.1 Assessing Security of Solution

The proposed approach meets *security and functionality framework* requirements. We first examine the intra-organizational version of the *security and functionality framework* since we are focused on privacy protection for the internal BQMA. Subsequently we discuss security within the inter-organizational context. We first discuss the security of exposing a field's characters. The above example will be assessed to show approach compliance. First, our approach is somewhat efficient. For a linkage field that needs to be character-analyzed, our approach stores the field's encryption and a list of ciphertexts. If n1 is the number of characters comprising the K_1 to be character-analyzed, the linked list contains n1 ciphertexts. An encryption of each field that will not be character-analyzed will also be stored for basic comparisons. Other approaches below entail more storage use, as well as other difficulties, as will be shown. Next, the security framework's "small secrets" tenet is supported. All the individual character plaintexts are securely encrypted, and the secret for decoding remains with the foundation algorithm's small encryption key. The proposed approach is public and not proprietary.

We discuss security against the AAE. We need to analyze the security of the foundation algorithm a deterministic encryption scheme. As mentioned in the *security and functionality framework*, full plaintext fields or individual characters should not be surmised due to the approach's transformations. Imagine that K_1 is broken down by characters, padded with K_2 in each record R_i. Imagine that N is the total number of records in a validation file. After deidentification, for each $R_i[K_1]$ we have new data structures $R_i[D]$, i=1…N. $R_i[D.id]$ and $R_i[D.hd]$ are specified. $R_i[K_2.id]$ has also been created.

First, notice that surmising individual characters by recognizing their distribution in various R_i would be unsuccessful. If through a *ciphertext-only attack*, for example, an employee were to observe only the ciphertexts, he could not surmise individual characters by recognizing the characters' frequencies. Suppose that K_1 is broken into characters. If E is the foundation algorithm, due to the K_2 padding, the frequency of $E(R_j[K_1][q])$ for some q across all j=1…N has no relation to the frequency of the underlying plaintext character $R_j[K_1][q]$ in the original linkage fields. The assumption is that K_2 is the same for identical units but different across different units across all their records, just like in the case of claims data representing different people, for example. The resulting $E(R_j[K_1][q])$ ciphertexts are much more distributed for all j=1…N. All the other AAE, beyond the *ciphertext-only attack*, can lead to the case where some number of R_i and their corresponding transformed privacy-protecting data structures become known. The other attacks signify that $R_i[K_1][q]$ and $E(R_i[K_1][q])$ for q = 1…n1 for some R_i have been discovered by an organizational employee. Without padding, this employee could easily find other j and r such that $E(R_j[K_1][r]) = E(R_i[K_1][q])$ for some R_j, i, and q, which were not part of the original data available via the AAE attacks. Other deterministically encrypted characters in other records will equal the encrypted positions of known R_i because, presumably, characters making up an identifier will be chosen from a relatively limited set of values. Padding stops such a vulnerability. The individual characters are the same for the same unit but different for all others, preventing any relationship to be surmised among the enciphered codes.[467]

[467] It's important to note that equal characters within a particular identifier might be surmised depending on actual distribution of characters within identifiers. As an illustration, if the 3rd and 4th characters of a given set of last names are always "ee," equal ciphertexts generated by deterministic encryption would be produced for these two letters within all these last names. The same corresponding padding is used for all

3.4.1.1 Exposure and Solutions for Using Identical Identifiers

There is more exposure of plaintexts, partial or full, involving fields belonging to the same units. Consider the discovery of other full fields belonging to the same unit. An employee can go through the rest of the validation file and find all other records belonging to the same units. Imagine k stands for K_1 or K_2, for illustration. If the plaintext for $R_i[k]$ becomes known for some i as part of the AAE, an employee can find other j such that $R_j[k.id] = R_i[k.id]$ since the same deterministic ciphertexts can be compared. For example, PM staff can even run the PM software and link all possible records for a given member id as the software will find equal member ids to link the claims records. These can be considered data compromises as such data were not uncovered earlier. Indeed, how does one perform record linkage while securing against the AAE? From a theoretical perspective, how can one determine equality of two messages when the system is subject to one of the more advanced attacks, such as at least a *known-plaintext attack*? Using a provided equality function, which can equate deidentified data, one can find all other equal messages in the data by simply calling this function to find the equal data. We should point out this is not just a problem of just our scheme but of any deterministic encryption scheme. Since one can compare the equality of deterministically-encrypted data one can find other equal records for exposed units.

To prevent such a disclosure, the system can prevent the employee from issuing direct commands to find specific identifiers. In the case of PM, PM staff could issue PM commands to find chronically ill patients; however, it could not specifically ask for all records associated with particular enciphered ids. Access to the sensitive $R_i[D.id]$ or $R_i[D.hd]$ structures could be placed behind a mechanism preventing the issuance of sensitive direct identifier comparisons or other sensitive queries. If the application staff doesn't need to see the identifiers but only to be assured that they are properly linked, the approach we describe should be appropriate. Behind a query restriction, our approach would link data for application operations. Further identifier access would not be needed if linkage were the main reason for the access.

records. Hence, depending on the broader values of last names within the domain of last names, it's possible to recognize that when the enciphered 3rd and 4th characters equal each other in certain last names, the 3rd and 4th positions of those last names must be the popular "ee." For example, if in a particular data set few other 3rd and 4th character positions equal, those that do will probably be the "ee" characters. Deterministic encryption preserves equal values. To address this problem, different padding could be used. For example, the 3rd character can always be padded with, for example, column 7 from the same validation record. The 4th character can always be padded with, for example, column 11 from the same validation record. Equality within identifiers would now fail because different padding is used for the 3rd and 4th characters. Equality across equivalent characters between records for the same person would succeed—subject to the caveat in the text that padding itself can also be in error, and thus undermine equality—because the same consistent padding has been used for those letters. Thus, linkage across records can be examined. The whole operation, including the comparison of enciphered letters, can also be placed behind an access control and new pseudonyms can be created for the application. We will discuss this latter approach further in the text as one approach for improving linkage security.

If some identifiers are necessary for operations, one method to provide access is to use new obfuscated identifiers. Such identifiers can be randomly assigned when two records are linked behind an access control mechanism. They would replace the $R_i[D]$ data structures as the visible identifiers available to the application.[468] When looking for specific units, the staff could supply such pseudonyms to locate those units' records. One way to create such identifiers would be to re-encrypt, i.e., use the encryption function again, on all validation file records. The value of linkage fields would not stay equal for long periods of time. A public key encryption system, such as one based on El Gamal, which allows for re-encryption, could be used.[469] Needed ciphertext, i.e., R[i][D] components, can be re-encrypted many times, each time producing successively unrelated ciphertexts. However, recovery of the original plaintext would still take only one decryption, which would be used for reidentification as needed by Enhancement 1. At a given frequency a background process could re-encrypt all the linkage identifiers, replacing each old one with a new one in its data set location. By the time an employee wants to learn more about the units he's found, the pseudonyms he supplies will have already been replaced and would not match the pseudonyms of the actual data. However, a shortcoming of creating new pseudonyms is worsened linkage. Unless intermediate obfuscated identifiers are kept when using this approach, it will be impossible to link data from a validation file to data from any prior validation file after the pseudonyms are recreated. The linkage fields would be different and direct equality comparisons of ciphertexts would be impossible.

Another way to address this problem is to recognize the particular needs of the application. It might be possible to construct a software "middle layer" between the application and data.[470] Such an application programming interface (API) or other software paradigm can provide suitable data access or answers to application computations, but would not allow for sensitive data access. This middle layer can be the work of future research and could be more flexible than the mere hiding of sensitive $R_i[D]$ data structures behind a query restriction, as above. As an example of this idea, in a database context, a SELECT statement with a WHERE clause specifying a single enciphered linkage identifier from a table column may not be allowed, which would address our dilemma above of how to do linkage yet not access enciphered identifiers. The "small secrets" tenet of the *security and functionality framework* would not be violated with such a "middle layer," however. Large quantities of deidentified, not identifiable, data are being placed behind for example an API. At worst, additional records of units already exposed via the AAE can be found. Data unrelated to the AAE-exposed data should remain secure as they are already deidentified.

When our approach is examined from an inter-organizational perspective, protecting from the above AAE exposure is hardly necessary. As long as the data are secure from a

[468] This is again specific to the organization's linkage data set implementation. In a database scheme, for instance, a new identifier column could be created.

[469] See "Special Topics in Cryptography: Electronic Voting: Why?" (class notes), 2004, <http://theory.lcs.mit.edu/classes/6.897/spring04/L17.pdf> (14 May 2005).

[470] See K.S. Candan, Sushil Jajodia, and V.S. Subrahmanian, "Secure Mediated Databases" (Proceedings—International Conference on Data Engineering, 1996), 28-9.

ciphertext-only attack they should remain secure as no additional practical threats could be possible. If record linkage is done inter-organizationally, as is more common for record linkage applications and data are exported by an organization to other entities for analysis, the ability to mount an AAE attack beyond a *ciphertext-only attack* is much less possible. The employees of the receiving entity should have minimal or no access to original plaintext data at the data-producing organization. They are not, presumably, employed by the data-producing organization. They can mount the *ciphertext-only attack* by, for example, trying to recognize original plaintext frequencies in the data. However, they cannot mount more advanced AAE attacks as they cannot see the original plaintexts or the encryption system which transformed them.

A *plaintext-brute force attack* is not feasible on our approach because the foundation algorithm was chosen to be secure against a *key-brute force attack.* Since the right key for our key-based scheme cannot be found, systematically going through all the plaintexts and generating the right ciphertexts is not possible.

3.4.2 Other Approaches to Protect Record Linkage

Other approaches cannot provide the needed security and functionality for Enhancement 1. We review several approaches to understand the challenges involved.[471] We examine approaches from an intra-organizational perspective as we are primarily focused on internal applications such as the BQMA. Less insecurity would be found examining the approaches from an inter-organizational perspective. Less protection is needed when attackers have minimal knowledge about an organization's internal data deidentification procedures. Our solution is more secure than other approaches in such a context, too, since we better protect against character frequency attacks, i.e., a type of *ciphertext-only attack.* Also, we will focus on solutions to the character-analysis dilemma. It is more challenging to solve since individual characters must be protected rather than fields in their entirety. If such a problem can be addressed, deidentifying fields in their entirety could be more easily addressed as well.

[471] We only discuss some possible approaches in the text. However, in addition to the approaches we discuss the following approaches have similar difficulties to those in the text: Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant, "Information Sharing Across Private Databases" (Association for Computing Machinery, Special Interest Group on Management of Data, June 9-12, 2003); Josh Cohen Benaloh, "Cryptographic Capsules: A Disjunctive Primitive for Interactive Protocols" (Proceedings on Advances in Cryptology – CRYPTO '86, Santa Barbara, California), 213-222; Ronald Fagin, Moni Naor, and Peter Winkler, "Comparing Information without Leaking It," *Communications of the ACM*, 39 (1996): 77-85; Ronald Rivest, L. Adleman, and M. L. Dertouzos, "On Data Banks and Privacy Homomorphisms," in *Foundations of Secure Computation*, ed. R.A. DeMillo, 169-177 (New York: Academic Press, 1978); G.R. Blakley and Catherine Meadows, "A Database Encryption Scheme which Allows the Computation of Statistics using Encrypted Data" (IEEE Symposium on Security and Privacy, April 22-24, 1985), 116-122; Catherine Quantin, François-André Allaert, and Liliane Dusserre, "Anonymous Statistical Methods versus Cryptographic Methods in Epidemiology," *International Journal of Medical Informatics*, 60 (2000): 177-83; Josep Domingo-Ferrer, "A Provably Secure Additive and Multiplicative Privacy Homomorphism," in *Lecture Notes in Computer Science 2433*, ed. AH Chan, 471-483 (London: Springer-Verlag, 2002).

Technical solutions to provide Enhancement 1 can be broadly classified as restricting access to data or changing the data themselves. Enhancement 1 can be classified as a secure function evaluation.[472] Linkage identifiers must be kept hidden. At the same time, employees must compute information based on individual characters to resolve linkage identifier errors. The literature demonstrates two possible solutions.[473] [474] [475] [476] [477] [478] [479] [480] [481] [482] [483] [484] [485] To keep data secret, one can restrict access to them. For example, access restriction systems include access control systems, which can limit access to data in databases and file systems.[486] [487] The data can also be modified to make them undistinguishable. For example, data perturbation approaches include methods such as adding noise. The point is presumably to make data less representative of actual facts.[488]

3.4.2.1 Access Restriction

Given the *security and functionality framework*, however, access restriction systems alone would be inappropriate for our character-level linkage paradigm. Such approaches are less secure per the *security and functionality framework.* Assume a function is written that does a needed string comparison computation behind an access control. A record linkage application requests computations over the original data without accessing the original data. The application may have sensitive linkage fields of many units to protect. For example, the PM application uses up to half a million, if not several million claims records, as discussed before. There would be a considerable amount of sensitive data to store behind the access control. Per the *security and functionality framework*, since

[472] Moni Naor and Kobbi Nissim, "Communication Preserving Protocols for Secure Function Evaluation," <http://www.wisdom.weizmann.ac.il/~kobbi/papers/sfe_proc.ps> (31 August 2005).

[473] Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection*, 63.

[474] Ravi Sandhu and Pierangela Samarati, "Access Control: Principles and Practice," *IEEE Communications Magazine*, 32 (1994): 44.

[475] C.J. Date, *Introduction to Database Systems, Sixth Edition* (Reading, MA: Addison-Wesley Publishing Company, 1995), 417.

[476] Russell.

[477] Shafi Goldwasser, "Lecture 7: Zero Knowledge" (handout given at lecture at MIT, August 2001).

[478] David Chaum, Claude Crepeau, and Ivan Damgard, "Multiparty Unconditionally Secure Protocols" (Proceedings of the 20th Symposium on the Theory of Computing, 1988).

[479] David Chaum, Ivan Damgard, and Jeroen van de Graaf, "Multiparty Computations Ensuring Privacy of Each Party's Input and Correctness of the Result," *Lecture Notes in Computer Science*, 293 (1988): 90-93.

[480] Zero Knowledge Systems, "Private Credentials," November 2000, 12-13, <http://osiris.978.org/~brianr/crypto-research/anon/www.freedom.net/products/whitepapers/credsnew.pdf> (31 August 2005).

[481] Goldwasser, *Lecture Notes on Cryptography*, 215-218.

[482] Catherine Quantin, "Anonymous Statistical Methods versus Cryptographic Methods in Epidemiology."

[483] Josep Domingo-Ferrer, "Advances in Inference Control in Statistical Databases: An Overview," <http://neon.vb.cbs.nl/casc/overview.pdf> (14 October 2003).

[484] Josep Domingo-Ferrer, "A Provably Secure Additive and Multiplicative Privacy Homomorphism."

[485] Dawn Xiaodong Song, David Wagner, and Adrian Perrig, "Practical Techniques for Searches on Encrypted Data" (IEEE Symposium on Security and Privacy, 2000).

[486] Date, 417.

[487] Russell.

[488] Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection*, 62.

smaller secrets are easier to protect than larger ones, such an approach, alone, would not be as secure. In our discussion of a "middle layer" between an employee and the sensitive data earlier, the idea was to place deidentified, not identifiable fields behind the "middle layer" to decrease potential for plaintext mishandling. We examine data perturbation solutions for Enhancement 1.

3.4.2.2 Non-cryptographic Data Perturbation

Data perturbation includes encryption and hashing and non-encryption and non-hashing techniques. Non-cryptographic techniques on their own would be inappropriate. For example, methods such as swapping or suppression may render data deidentified.[489] However, such methods may also undermine record linkage. Linkage fields would become perturbed, or removed, in case of suppression.[490] Linking records becomes much more difficult, if not impossible, as a consistent field is no longer available. Although encryption and hashing techniques appear appropriate, existing work does not appear useful. Consider relevant deterministic schemes that transform the same plaintexts into the same ciphertexts. We examine several schemes.

3.4.2.3 Summary of Our Scheme

We should remember, our approach to securely address linkage identifier errors encrypts the individual characters to be character-analyzed after concatenating them with padding, which is another field from the same validation record. As long as the padding is error-free, any algorithm which compares the equality of individual characters in the compared fields can be implemented using our scheme. A robust string comparator might be utilized with our scheme to, for example, compute the *distance* between two fields in spite of deidentification. Equality of enciphered corresponding or non-corresponding characters between two fields may be compared since the enciphered characters should preserve any original plaintext distance relationship. We compare such capability with the similar capability and privacy protection of other approaches.

3.4.2.4 Encrypted Search with Identifier Errors

Song offers an approach for untrusted users to find words on behalf of trusted users in encrypted documents.[491] However, her idea is inappropriate due to the need to enumerate errors, which may not be readily done. The basic approach is a comparison of

[489] See for example Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection*, 62.

[490] See Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection*, 58.

[491] Dawn Xiaodong Song.

deterministically encrypted text. A deterministically encrypted word is compared to a deterministically encrypted word combined with a pseudo-random (random-like) word in a document containing many such words, i.e., each encrypted with a different pseudo-random word. To find a word in the document, the desired word is first deterministically encrypted. If this word and a particular word in the document equal after basic deterministic transformations, the plaintexts equal, and the position in the document may be returned to a requesting user. The location of the requested word has been found. In the record linkage context, such a scheme could be simplified to a comparison of deterministically encrypted identifiers. It can be the case of seeing if a word exists in a "document" which is itself only one word long. If ciphertexts equal the plaintexts equal due to the deterministic transformations. Song proposes that *regular expressions* be used for error handling. When examining deterministically encrypted words, specific character positions in the word can be generically specified to allow for various words matching those positions to be compared. The system encrypts all the possible resulting words for comparison. For example, comparing the word "ab[a-z]" may generate a list of 26 enciphered words: "aba" through "abz"—each would be compared to words combined with pseudo-random words in a document, as described above. However, such an approach would not be useful for our context. The approach suggests that the user may know which errors might be possible. Such knowledge would feed the encryption algorithm. A given field would be converted into several versions of itself to cover several possible errors for comparison. This greatly limits the approach as users may not know all relevant errors. They might have general perceptions of error types, they might even examine the training file, but they may not be aware of the full range of errors which can arise within a validation file due to prior processing, error cleaning, etc. Matching performance might decline.

3.4.2.5 Matrix of String Comparison Metrics

Du presents ideas which appear more appropriate for our context.[492] The intent is to pre-compute all comparison results between deidentified linkage identifiers instead of pre-computing just some results as Song's error-handling ideas suggest. However, these ideas are also inappropriate because under a BQMA context, using Du's scheme would entail unrealizable storage and could lead to discovery of considerable plaintext in dense space. Du attempts to securely calculate pattern matching scores between a query and a database of records just as in our string comparator discussions. Various scenarios are presented depending on who owns the database, which of several parties performs the query, and from whom information must be hidden, for example, from the owner of the database or from an intermediary helping perform the query. Assume b is an n-character length query and t is one of the n-character length strings in the database. A pattern matching score S is to be computed:

[492] Wenliang Du and Mikhail Atallah, "Protocols for Secure Remote Database Access with Approximate Matching" (7th ACM Conference on Computer and Communications Security (ACMCCS 2000), The First Workshop on Security and Privacy in E-Commerce): 1-25.

$$S = \sum_{k=1}^{n} f(b_k, t_k)$$

The function f is a basic metric between two strings, such as functions like $|b_k - t_k|$, $(b_k - t_k)^2$, etc., similar to our string comparisons from before.[493] To compute score S securely, Du suggests that the values S for b and every t in the database be calculated ahead of time.[494] Rather than computing such a score in real-time, the score is pre-computed to be referenced later. We'll consider one implementation to understand the challenges involved. Applying such an approach to record linkage, since this approach compares two linkage fields, any one of them can be the "b" or the "t" as needed. A large matrix structure can be created allowing all linkage fields to act as the needed "b" or "t" identifiers. We first discuss the case when all possible units are included. We will investigate below what happens when we pre-compute only some of the linkage field combinations based on units that actually exist in the validation file, which might be more secure. Also, instead of computing just a summation of individual-letter metrics across two strings, we can compute a single score as a function of both full strings. This should offer more flexibility for error-handling as full strings can be accessed and non-corresponding characters can be examined.

The created matrix will contain scores S. A given score will represent the comparison of a linkage field in that cell's row to a linkage field in that cell's column. A pseudonym can be assigned to each linkage field for reference purposes. Any secure encryption scheme—indeed, any consistent secure string transformation—can be used to create the pseudonyms. The pseudonyms would be assigned to the row and column positions in the matrix corresponding to the linkage fields they represent, and they would replace the fields in the actual records in the validation file to do the field comparisons. During field comparison, the application can obtain the two pseudonyms from the two compared file records, check the row of one and the column of the other in the matrix, and find the appropriate score S to determine those linkage fields' equality. Such an approach will not work in our context, however. First, it is very inefficient. To create the matrix itself, space would have to be allocated for M * M pre-computations, where M is the total number of values in a linkage field namespace for a unit. In the 0000000000 – 9999999999 domain of member ids for PM, as before, room for $10{,}000{,}000{,}000^2$ scores S would have to be allotted. This would involve unrealistic storage given today's computational capabilities.

Also the approach is less secure than our proposal of securely encrypting a linkage variable's individual characters. An important question is how to create the matrix so that assignment of pseudonyms to file records is straightforward. Record linkage fields can be easily transformed before they enter a validation file. One way is to create the matrix with each column and row representing a systematic increase in linkage fields. For example, in the case of PM, the columns of the matrix can systematically increase from 0000000000 to 9999999999, for example, from left to right. Each column would hold the place for the member id of that value and contain a randomly generated pseudonym which would

[493] Du, 17.

[494] Du, 15.

represent the member id. Similarly, rows can systematically increase in the same range from top to bottom. Each row would hold the place for the member id of that value and contain a pseudonym equal to the pseudonym of the same *column* number because both row and column must represent the same member id. Assigning the proper randomly generated pseudonym to a linkage field for an incoming validation file record is easy. The staff/process generates the pseudonym for the linkage identifier, finds the linkage field's ordered row and column positions in the matrix, places the pseudonym into these two locations, and switches the linkage field in the file record to the same pseudonym. However, when the matrix is published for the application to perform variable comparisons using pseudonyms, the employee operating the application and any others would immediately know exactly what is the plaintext corresponding to the pseudonym for a given linkage field in a file record. Since the matrix is ordered from top to bottom and left to right, an employee can recognize the original plaintext value. The distance between the beginning of the plaintext namespace to the identifier value is the same as the distance between the beginning of the matrix and the pseudonym's row or column position. Thus, the original plaintext value is the difference between the beginning of the matrix and the pseudonym's row or column position, added to the beginning of the plaintext identifier namespace.

The columns and rows can be randomized. Random columns can be switched with each other, and random rows can be switched with each other, all while preserving the pattern matching score S in each matrix cell. Such switching would prevent the "decryption" attack just described. In addition, only the "active" columns and rows can be released to the application, not the entire matrix. Staff will have access to rows and columns only for the units for which records are available. In the PM analogue, member ids of policy-holders who actually used the health care system in a given past period, a period specified by the PM application, but often a rolling 6 or 12 month claims period, would be the ones whose member ids and scores are included. Scores for non-users would be absent. This should further limit any use of the matrix to decrypt pseudonyms as many intermediate pseudonyms used for "decryption," i.e., guiding, comparative purposes would not be available.

However, neither approach will suffice. Sometimes the matrix is relatively full. For example, the copy data store may be dense and diverse regarding member ids in a typical 12 month period, as was one of our assumptions before. For a given linkage field, many if not most of the values in its namespace should be present in the matrix, despite attempts to publish only *active* linkage fields. Combining this fact with any of the *attacks against encryption* (AAE) at least as strong as a *known-plaintext attack*, and, indeed, the pattern matching scores available in the matrix, might allow for an employee to decrypt a number of pseudonyms beyond those available to her via the AAE. Staff can plug in the plaintexts it knows of and other alphanumerically close linkage variables into function f, the basic comparison between two strings. Function f should be public, otherwise, Du's approach would be less secure per the *security and functionality framework* as f is proprietary. Staff can generate scores S and compare them to those in the matrix under the column or row positions of the pseudonyms it has available via the AAE. For each S, staff may find several or possibly even one matrix cell which contains the same score(s).

The pseudonym(s) associated with that row or column must correspond to the plaintexts for which score S was just computed. Since staff knows the linkage fields it supplied for the f computation, a pseudonym(s) close to one of the pseudonyms uncovered via the AAE has just been uncovered, as its underlying linkage identifier(s) was used for the f computation. This can happen for all the plaintext-ciphertext pairs available to an employee via the AAE, exposing considerable plaintext.

3.4.2.6 Hashed Linkage Field Characters

Churches' method comes closer than the above approaches.[495] The idea is to use hashed bigrams of the linkage identifiers to compare the identifiers' similarity in real-time instead of pre-computing the comparisons. However, his ideas are still less appropriate than our approach as the ability to recover a number of smaller units of plaintexts is a key weakness. Churches tries to implement a "blindfolded" record linkage process similar in spirit to the approach of this thesis. An analyst must link two data sets without knowing the values of the linkage variables. Each record in each data set has multiple fields on which record linkage can be performed. When comparing two analogous fields X and Y, one from each record, a *dice coefficient* value, a string comparator value, can be calculated measuring the two fields' similarity:[496]

Dice coefficient = 2 * |bigrams(X) ∩ bigrams(Y)| / (|bigrams(X)| + |bigrams(Y)|)

The function bigrams(X) takes word X and breaks it into its bigrams, which are all the overlapping subwords of X of length 2. Record linkage is done as before. During system training, the dice coefficient across all fields to be matched is computed. The resulting values feed directly into a weighting process. For example, the agreement and disagreement weights for equality and inequality patterns can be computed. During matching, the sum of the weights is compared to a threshold which signals if the two records represent a match. To preserve the privacy of such an approach, Churches creates a data structure which is substituted for each linkage identifier in each record just as in our proposed approach. First, Churches creates a power set of each bigram set, i.e., an exhaustive set of subsets of bigrams(X) for each field X to be linked. Next, Churches hashes each such subset with a Hashed Message Authentication Code (HMAC), essentially a keyed hash function. Using an HMAC prevents a *plaintext-brute force attack*.[497] Systemically hashing all possible plaintexts to find the hash codes available in the system is not possible without knowing the key, which presumably would not be available to staff. The hashed results, the length of the hashed subset, and the length of the original number of bigrams are all inserted into a new privacy-protective data

[495] Tim Churches and Peter Christen, "Some Methods for Blindfolded Record Linkage," *BMC Medical Informatics and Decision Making*, 4 (2004): 1-17, <http://www.biomedcentral.com/content/pdf/1472-6947-4-9.pdf> (24 May 2005).
[496] Sam's String Metrics, "Dice's Coefficient," <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#dice> (30 May 2005).
[497] Churches, 4.

structure. For example, the following table, taken from Churches' article, represents elements of such a structure when the linkage field to be used for comparison has value "peter."[498]

[498] Churches, 6.

List of bigram subsets, their hashes, sizes of subsets, and size of bigram('peter')

A record key	*A.a bigram subset*	*A.a_bigram_combination_digest*	*A.a_bigram_ combination_length*	*A.a_length*
10	('er')	0a3be282870998fe7332ae0fecff68cc0d370152	1	4
10	('et')	8898f53d6225f464bb2640779cb17b9378237149	1	4
10	('pe')	6fc83a87ee04335a58aa576cb5157625b1b5c51b	1	4
10	('te')	f2bcfb3d76d7fc010e3adc08663090f29c5e928a	1	4
10	('er', 'et')	f86abb0c84889d004b817e86199b3837708d70e9	2	4
10	('er', 'pe')	df99d8658d8165af4552f60ade3662ba98006298	2	4
10	('er', 'te')	edfb618d37ecfafc9735e6ad4675245a4071aa9d	2	4
10	('et', 'pe')	bd7ada000c2b9004b7519b989bfcfdff7ad36678	2	4
10	('et', 'te')	fdcb71db96d2da9b1d19b62944c5f36448cb2668	2	4
10	('pe', 'te')	71322eeebabff9828aeed3281a86577163e16a78	2	4
10	('er', 'et', 'pe')	8bf2788ef28443b7a0298f19defa5532db40f63a	3	4
10	('er', 'et', 'te')	c7e9a32e54ba33d3769c4813616fdfcc6306459c	3	4
10	('er', 'pe', 'te')	33287ce86aa02af0f31d4857a79671c1f4645277	3	4
10	**('et', 'pe', 'te')**	**ecd7b151291f1612595c9f8f385e9f71119a1ae0**	**3**	**4**
10	('er', 'et', 'pe', 'te')	65e568493a08a3428595b8be35f6ae2a0f48d170	4	4

The second column in this data structure represents the power set of the bigram set, i.e., the exhaustive set of subsets of bigrams(X). The third column contains the corresponding hashed values of these subsets. The fourth column is the size of the subset investigated in column two. The fifth column is the size of the original bigrams(X) result. In the case of "peter" there are four total bigrams, as shown at the bottom of column two. The third, fourth and fifth columns enter Churches' privacy-protective data structure.

When comparing two validation files, the records of which are to be matched (or one file, if linking a file to itself), an analyst performs an inner join of the third column in the table across all the corresponding fields in both files to form a cross-product. All matching hashed values from the data structures representing particular record fields in the two different data sets to be linked are found. For each row in the inner join, the analyst computes the dice coefficient. Imagine using the above dice coefficient formula and the above table columns which are now in the privacy-protective data structures. For field X in one validation file and corresponding field Y in the other file, the above dice coefficient formula becomes:

2 * the fourth column of X / (the fifth column of X + the fifth column of Y)

Based on such a computation the system can assess match status. The highest dice coefficient value is selected for each unique pair of fields X and Y across both validation files. This value feeds the weighting and matching process as described in our record linkage discussion in Section 3.2.2.2. If two records produce a total weight, comprised, in part, of the weights based on the calculated dice coefficient values across the compared fields for which the dice coefficient was computed, at least as high as the threshold, the records match.

Such an approach would be inappropriate for our context. First, the use of hash functions is inappropriate. Hashing is an irreversible process. As referenced before, it's not possible to "de-hash" the ciphertexts and obtain the original variables. Reidentifying data subjects would not be possible. Of course, staff can get the HMAC key and go through the entire list of possible linkage fields, hashing each one to find the enciphered linkage fields in their possession (essentially a *plaintext-brute force attack* on the results generated by the application). The original fields could be found when the hash codes generated equaled those available. This is a less efficient approach.

However, a larger problem is the use of bigrams. Due to the more advanced AAE, if some plaintext linkage fields and their corresponding ciphertexts become known, considerably more information can be learned about plaintexts of linkage fields unrelated to those discovered. Assume that the plaintext for some single bigrams and their hashes become known. Presumably a number of identical characters exist in other linkage fields due to character homogeneity in the domain of a linkage identifier. Namespaces should not be created anew for each identifier. Therefore, other equal 2-character plaintexts could be found by comparing equal ciphertexts available via the AAE. In Churches approach, therefore, a number of equal 2-character plaintexts for units could be discovered, unrelated to those of units already found.

3.4.2.7 Secure Function Computation

Feige presents a general way to compute any function.[499] However, Feige's approach is also inappropriate for our context as impractical amounts of storage are involved.

[499] Uri Feige, Joe Kilian, and Moni Naor, "A Minimal Model for Secure Computation," 1-15, <http://www.wisdom.weizmann.ac.il/~naor/PAPERS/fkn.pdf> (24 May 2005).

Specifically, Alice and Bob, possessing inputs a and b respectively, preprocess them and give the results to third party Carol. Carol will compute F(a,b), yet she should not be able to surmise a or b. We can explore Feige's case for F(a,b) = 0 or 1, i.e., a Boolean function. Equality can be considered Boolean. Linkage identifiers either equal or do not. However, when they have errors, equality may not be Boolean. A metric, such as from a Levenshtein string comparator, might be assigned to indicate degree of equality. We'll explore the Boolean case to understand the challenges involved. We should point out, however, that this three-party paradigm differs from many other secure function evaluation approaches in the cryptographic literature. They typically focus only on two parties, e.g., Alice and Bob, performing local computations on their inputs and communicating to compute a given function without one party learning of the inputs of the other. Such an interactive two party framework would be inappropriate for our record linkage paradigm as there is no "interaction" within our paradigm, and the analyst running the software should not know of any of the inputs. Deidentification may happen in another department within the same organization or in two different organizations if data are to be obtained from multiple sources for centralized analysis. The process obfuscating one linkage identifier value has no need to "hide" that value from another process obfuscating another field value as they are only trying to protect against a downstream user. Also, the record linkage analyst will only have access to masked data, and, by intent, should not know any of the original plaintext data.

In Feige's approach, Carol computes the needed F(a,b) result by processing a list of pre-computed intermediate results created by Alice and Bob. First, Alice and Bob pick common random bits, which will hide their values a and b, respectively. Alice pre-computes F(a,b) for every b in the domain. In particular, she uses some of the random bits selected with Bob to permute the order of her F(a,b)'s. To each F(a,b) result she XORs one of the remaining random bits. Both of these transformations will be reversed by Carol. At the same time, Bob uses some of the same random bits to hide his b by permuting its structure. This b will become an index within Alice's list, used to locate the F(a,b) results. Alice sends her permuted list to Carol, and Bob sends his permuted b to Carol. Also, Bob sends the appropriate random bit to Carol to ultimately de-randomize the F(a,b) value. To compute F(a,b), the created permutation and randomization must be reversed. Carol selects the proper F(a,b) from Alice's list based on Bob's permuted index; she follows the pointer b to get an intermediate F(a,b) value which reverses the permutation Alice originally created. To get the final F(a,b) value, Carol decrypts the intermediate value and XOR's it with Bob's random bit, i.e., de-randomizing the value.

To make this approach work in our context, a linkage field's permutation can become its pseudonym. This pseudonym could be used for comparison: when this linkage variable is compared with another, the comparison result could be found by using the linkage variable pseudonym as an index into the other linkage variable's comparison list to find the intermediate result. The de-randomization bit could be stored with every linkage field and applied to this result to get the final comparison result between two fields, i.e., a Boolean 0 or 1.

However, such an approach would not be applicable to our context because it requires unrealizable storage. First, to make it workable, each comparison list and each linkage field would need the same permutations and randomization bits. Otherwise, an unrealistic amount of storage would be required. First, assume the approach implements the same permutations and randomization bits. The proposal would still be inefficient. Each identifier must contain a permuted list L to compare that identifier with all others in the domain to get final F(a,b) results. Each identifier in the domain could be represented by the index b, pointing to a particular position in L, and the de-randomizing bit to get the Boolean "F(a,b)" result representing the equality comparison between that identifier and the L identifier. If PM member id values are 0000000000 – 9999999999, as before, this namespace could be represented by at most 34 bits. Feige's approach ultimately operates at the bit level, therefore, 2^{34} values would need to be allocated to represent this 10-billion identifier domain. Two lists each with 2^{34} items would have to be constructed. The first list will be all the L identifiers permuted in the same way, each representing a different member id. The second list will contain tuples representing all the member ids represented as index b and the particular derandomization bit based on a particular set of randomization bits. To compute final F(a,b) values will thus require $2 * 2^{34}$ or 2^{35} stored items. A given member id will point to a particular place within an L identifier. The associated derandomization bit will be applied to this found result to get the final F(a,b) value. Having 2^{35} stored items would be unrealizable given today's computational capacity. Having different permutations and sets of randomization bits to randomize the member id namespace in different ways would require even more storage. In this case, 2^{35} billion new items would have to be constructed for *each* unique combination of permutation and randomization bits, because in addition to the lists as described above data would have to be stored indicating which permutation and randomization bits a particular comparison requires. Such storage requirements may be even less practical. Approaches involving more complicated F(a,b) calculations beyond Boolean would also involve more storage. Simplifying this approach and having, for example, a single matrix with pre-computed results of every linkage variable compared with every other is a possibility. The comparison protocol is simplified. There are problems with such an approach, too, as was discussed regarding Du's article above.

3.5 Predictive Modeling Experiment

To demonstrate the practicality of our Safe Harbor deidentification and identifier error mitigation approaches we conducted an experiment with a PM application as used by insurance organizations. Our results suggest that privacy protection can be practically implemented within the BQMA. We licensed a PM application from DxCG, a company that creates predictive models for a variety of health industry stakeholders.[500] [501] DxCG clients, such as health plans and providers, currently cover 62 million lives. The DxCG PM examines patient diagnoses and basic demographic data to predict the medical risk

[500] DxCG, <http://www.dxcg.com/> (1 September 2005).

[501] DxCG, *About the Company*, <http://www.dxcg.com/about/index.html> (3 April 2005).

and financial cost of managing such patients from a health insurance perspective. Our goal was to compare the DxCG PM performance using identifiable versus deidentified data, the latter created as per the Safe Harbor principle. We obtained billing data from the Beth Israel Deaconess Medical Center (BIDMC), a hospital in Boston, MA. This billing data would be used to construct the UB92 records that a provider would submit to insurance organizations, which in turn would use the data to identify and risk stratify chronically ill and high-risk individuals using PM. The BIDMC Institutional Review Board approved our usage of the data. The DxCG PM RiskSmart version 2.0.2 was used for our analysis. We used the MySQL 4.1 open source database to store the data, link them, and deidentify them for DxCG PM operations.[502]

We used DxCG PM's ability to predict future financial cost of patients, modeling a real disease management application as used by insurance organizations.[503] [504] We used the DxCG PM to predict costs for the subset of 32,294 patients in 2004, for which demographic data were available, who were hospitalized in 2003.[505] In 2003, 12,754 of these 32,294 individuals had a total of 258,570 hospitalization diagnoses with a number of people having multiple diagnoses. We focused on DxCG PM's ability to predict 2003 individuals who would cost at least $25,000 in 2004, focusing on high-cost ("high-risk") patients.[506] [507] The 2004 expenses predicted by the model for the 32,294 individuals were compared to the actual expenses BIDMC incurred in 2004. The model's predictive performance was quantified using positive predictive value (PPV) and sensitivity.

Enhancements 1 through 6, from Section 3.1.1, can impact an application's performance in two ways. They can undermine application performance because data fields are removed or modified, limiting computation. They can improve application performance because despite record linkage variables being deidentified, data can be better linked which can improve analysis. The intent of our experiment was to isolate the impact of both effects to test each.

We tested the performance of the DxCG PM on fully identifiable data, a *control* scenario. PPV and sensitivity were 43% and 35%, respectively.[508]

[502] MySQL, <http://www.mysql.com/> (1 September 2005).

[503] DxCG, *RiskSmart Models and Methodologies* (2002), 5-15.

[504] Ingenix Corporation, "Identification and Management of High Risk Patients Using a Claims-based Predictive Model," <http://www.ingenix.com/esg/resourceCenter/10/~dt_pyr_wp_1-03.pdf> (1 September 2005).

[505] Throughout this experiment we configured the DxCG PM with the following standard configuration parameters: commercial population; inpatient model; explanation model; prospective model; model predicts medical expenses only without truncation. (See DxCG, *DxCG RiskSmart Stand Alone User Guide Version 2.0.1*, 8). Also, as will be seen in the text, the DxCG PM allows one to specify how many months during the year a patient was eligible for health insurance to properly predict risk. This value was allowed to default to 12 in our case, a generic value, as the Beth Israel Deaconess Medical Center, a hospital, does not have insurance eligibility information.

[506] DxCG, *RiskSmart Models and Methodologies*, 5-15.

[507] Ingenix Corporation, "Identification and Management of High Risk Patients Using a Claims-based Predictive Model."

[508] In discussing our experiment with DxCG staff to ensure proper DxCG PM use, a DxCG staff member recommended the DxCG PM output be "normalized" to our population. (Katherine Salerno, DxCG staff,

3.5.1 Deidentifying Claims Data

We deidentified the data.[509] To understand how deidentification impacted the data, we describe the data structures and data relationships in the DxCG PM. The DxCG PM relies on patient diagnoses, date of birth, gender, and eligibility months for prediction purposes. These items are spread between the DxCG PM's diagnosis file and eligibility file. The diagnoses and medical record numbers are in the diagnosis file, representing individuals with particular conditions. The eligibility file contains the eligibility months, i.e., the number of months individuals were eligible for health insurance; genders; birthdates; and medical record numbers associated with the individuals. Independent risk factors for gender, age, and all of a person's diagnoses produce separate risk scores which are added together and combined with the number of months a person was eligible to produce a person's total financial projected cost.[510]

Eligibility months or gender were not changed per Enhancements 1 through 6, which allow these values to remain unaltered.

Diagnoses were slightly changed based on the requirement to remove "any other unique identifying number, characteristic, or code," point 14 in the Safe Harbor list from before. It is unclear what exactly is meant by "unique identifying" in the HIPAA language as the law is still relatively new. However, one possible explanation is that the prevalence of the condition represented by the diagnosis must be very low in a local area or nationwide.[511] [512] We ran a histogram on the 1,546,963 diagnoses in the full file which was used in the DxCG PM of which the 258,570 diagnoses were utilized for our experiment. Approximately 850 diagnoses were unique. Treating such codes as "unique" per Safe

email to author on September 29, 2005). The BIDMC data showed that the hospital's patients were more ill than the population on which the DxCG PM was calibrated, requiring an inflation to the program's risk scores to properly predict risk. Throughout our experiment, the following computations were done on the output from every DxCG PM run: 1) the "prospective relative risk score" for each patient, i.e., the predicted risk score generated by the software, was divided by the average prospective relative risk score for our population of 32,294 patients; 2) the resulting risk score for every patient was multiplied by the average actual future (i.e., 2004) costs in our population. These "normalized" costs became the predicted costs on which PPV and sensitivity for the DxCG PM were computed in our analysis.

[509] Besides Enhancements 1 through 5, no further deidentification was done as a result of "actual knowledge" analysis, Enhancement 6. The BIDMC apparently does not require further data deidentification. (Implied, Meghan Dierks, M.D, Instructor in Medicine, Harvard Medical School, personal interview, July 21, 2005). The Massachusetts Institute of Technology also does not require additional deidentification beyond the removal of the Safe Harbor items mentioned in Section 3.1.1 in the text. (Massachusetts Institute of Technology, "De-identified Data," *Committee on the Use of Humans as Experimental Subjects*, 30 November 2004, <http://web.mit.edu/committees/couhes/definitions.shtml#De-identifiedData> (1 September 2005)). As we are following common institutional practices, we also did not analyze or remove any data further. Future change in practice might change this analysis.

[510] DxCG, *RiskSmart Models and Methodologies*, 5-6.

[511] Shannon Hartsfield, attorney and HIPAA specialist, telephone interview with author, February 26, 2004.

[512] Taken from John Neiditz, attorney and HIPAA specialist, telephone interview with author, February 26, 2004.

Harbor, because they were unduplicated, we shortened all these diagnoses to 3 digits from 5, generalizing the diagnoses to more common illnesses. We ran another histogram, and the diagnoses that continued to be unduplicated were deleted.

Date of birth (DOB) was changed to age, as allowed by Enhancement 3. Enhancement 3 also requires people over 90 to be aggregated into a single age category. Individuals over 90 were relabeled to 90 years of age.

The deidentified data produced the same PPV and sensitivity. The diagnoses changes did not change PM performance since other diagnoses for individuals in the eligibility file were sufficient to properly categorize individuals' risk. The DxCG PM could use age instead of DOB as the model could be specified to work with age as an input parameter. Within our population, the future cost of people over 90 years of age did not apparently materially differ from those of age 90.

Some PM applications rely only on diagnosis, number of months covered by the health insurer, age, and gender for prediction purposes, suggesting that a number of PM applications in US health insurance organizations may work in a deidentified fashion.[513] [514] [515]

3.5.2 Improving Linkage of Claims Data

Allowing for deidentified variables to fix record linkage errors improved the DxCG PM's performance as data synthesis improved. To control the linkage process, two linkage variables were created. A linkage field, K_1, was created from a consistent patient identifier in the BIDMC data with the following properties:

Number of character positions in $K_1 = 10$
Range of each character position = 10 (i.e., each character could be digits 0-9)

A second linkage variable, K_2, was created from the same consistent identifier from the BIDMC data with the following properties:

Number of character positions in $K_2 = 12$
Range of each character position = 10

As indicated before, we assume a uniform distribution of all values in the namespace for K_1 and K_2 for simpler presentation. Our analysis can be carried out with the non-uniform namespace constructions more typically found in practice. We first tested a system with fully identifiable data using only K_1 for linkage. PM performance worsened when we

[513] DxCG, *DxCG RiskSmart Stand Alone User Guide Version 2.0.1*, 24-29.

[514] Symmetry, 5-6.

[515] Arlene Ash, Yang Zhao, Randall Ellis, and Marilyn Schlein Kramer, "Finding Future High-cost Cases: Comparing Prior Cost versus Diagnosis-based Methods," *Health Services Research*, 36 (2001): 195.

introduced errors into the linkage variable. We randomly altered K_1 for roughly 25% of the 258,570 diagnosis records for the 12,754 year 2003 patients in the diagnosis file. PPV of the model slipped to 41% while sensitivity remained at 35%. The diagnoses for some individuals had medical record numbers in the diagnosis file which no longer mapped to the medical record numbers of particular individuals in the eligibility file. In other cases, new diagnoses were created for existing people by transposing the medical record numbers of unrelated individuals into *their* medical record numbers. Some people appeared to acquire new health status.

Using a better K_2 enhanced record linkage. We introduced a 5% error into K_2 before data deidentification. Afterwards we encrypted K_2 with a deterministic method and deidentified the rest of the data as before.

K_2 was a useful field for the system. Despite deidentification, we could still use our record linkage techniques from earlier. Comparing field length and error rates, which can be provided to PM staff by the staff or process deidentifying the K_1 and K_2 earlier, PM staff can find another optimal field for linkage. It can compute equation (17) from Section 3.2.2.3.3 and determine whether a new field would fix the mistakes of existing linkage variables. In the case of K_2, (17) was satisfied:
$(p_1^{n1})*(1 - 2*e)*(2*f) < (2*e)*(p_2^{n2})*(1 - 2*f)$
➔$(10^{10})*(1-2*0.25)*(2*0.05) < (2*0.25)*(10^{12})*(1-2*0.05)$
➔$5.0*10E+8 < 4.5*10E+11$

Again, we stress these computations are based on the assumption of the uniformity of the K_1 and K_2 namespace distributions. Since our theory can be converted to work with the more common non-uniform distributions in practice, a similar analysis to the one here can be constructed. The deidentified K_2 also introduced fewer overall errors into the system, meeting (19):
$(1/p_2)^{n2} + 2*f < 2*e + (1/p_1)^{n1}$
➔$(1/10)^{12} + 2*(0.05) < 2*(0.25) + (1/10)^{10}$
➔$0.1 + 1.0*10E-12 < 0.5 + 1.0*10E-10$

When linking the data using the deidentified K_2 and K_1 some errors created when only the plaintext K_1 was used were fixed by the K_2 field. PPV rose to 42% while sensitivity remained at 35%.

3.5.3 Additional Techniques for Claims Deidentification

As indicated in Enhancements 1 through 6, Safe Harbor also requires deidentifying several other items in a UB92 not used by the DxCG PM, including:

- Patient Address. The zip code can remain, but can be no longer than 3 digits.
- Patient Name, Insured Name.

- Several dates: Admission Date, Occurrence Span "From" Date, Occurrence Span "Through" Date, and other dates. Only the year can remain for such dates.
- Treatment Authorization Codes. These should be unique for each patient and therefore must be removed.
- Employer Name, Employer Location. These must be removed per the Office for Civil Rights, as referenced earlier.[516]
- Provider information such as Provider Number or Other Physician Ids (up to 2 of them).
- Remarks. This free-form text field must be removed as it may contain unique information that may help identify individuals.
- Other fields with very unique values or claim records with combinations of deidentified fields that are unique in the claims data. The former types of fields must be removed. In the latter case, the entire claim record must be deleted, some of its fields must be generalized, or other deidentification techniques be applied as policy-holders might be identifiable based on "actual knowledge" analysis due to unique data combinations. Unique records within a file might be linked to other data sources which can help reidentify individuals.

However, the original unaltered fields may be usable by a PM application:

- A full zip code can be used.[517] For example, a zip code may imply a higher probability of acquiring particular chronic conditions.[518]
- Removing an individual's name would prevent the recognition of her ethnicity or race, which may be needed by PM to characterize risk.[519] [520] [521]
- Removing day and month in dates would undermine detailed date analysis, which may be needed by PM. For example, some applications compute hospital length of stay; recognize a patient's accelerating utilization of health services; or try to understand the patient's prescription compliance pattern to better identify high-

[516] US Department of Health and Human Services, Office for Civil Rights, "Standards for Privacy of Individually Identifiable Health Information."

[517] Laura Benko, "Long-range Forecast: Partly Healthy, Chance of Storms," 26.

[518] Tatlow showed that the geographic density of alcohol outlets within a zip code was a significant predictor of alcohol-related hospital admissions. (See James Tatlow, John D. Clapp, and Melinda M. Hohman, "The Relationship between the Geographic Density of Alcohol Outlets and Alcohol-related Hospital Admissions in San Diego County," *Journal of Community Health*, 25 (2000): 79). Alcoholism might be considered a chronic condition as it often can only be controlled and not cured. (Lander University, class notes for NURS 416, <http://www.lander.edu/bfreese/416%20Notes%20Ch%2021.doc> (14 October 2003)).

[519] Kiran Nanchahal, Punam Mangtani, Mark Alston, and Isabel dos Santos Silva, "Development and Validation of a Computerized South Asian Names and Group Recognition Algorithm (SANGRA) for Use in British Health-related Studies," *Journal of Public Health Medicine*, 23 (2001): 279.

[520] Iezzoni, 45.

[521] Joe V. Selby, Andrew J. Karter, Lynn M. Ackerson, Assiamira Ferrara, and Jennifer Liu, "Developing a Prediction Rule from Automated Clinical Databases to Identify High-risk Patients in a Large Population with Diabetes," *Diabetes Care*, 24 (2001): 1550.

risk individuals.[522] [523] [524] [525] [526] In addition, full dates of birth and ages beyond 90 as deidentified by Safe Harbor may also be predictive of patient cost.[527] [528] For example, a newborn's date of birth can identify the season when he was born, which may make him more susceptible to certain health problems.[529] Perls found people over 90, unexpectedly, might have health care costs lower than 65-90 year old individuals.[530] Perls showed that 90+ individuals might have already experienced a particular condition, and thus don't need as intensive treatment as those 65-90 years old because the condition is not affecting them for the first time, as well as other reasons, why their care costs might be lower than for individuals 65-90 years old.[531]

- Care authorization information can be used to predict care utilization.[532] Liu showed how number of visits authorized to a mental health provider by a managed care mental health organization was less than the number of visits that could have been desired by the provider. However, providers did not renew patient authorizations` and thus did not offer patients maximal possible treatments. They may have perceived administrative difficulties in seeking reauthorizations, or were concerned that the mental health organization might view them less favorably compared to others within the provider network who treated patients with fewer visits.[533]
- Employer information may be predictive of high risk as some occupational environments create increased risk for certain health maladies.[534]
- The nature of a provider's practice may be used to assess patient care quality implications inherent in the practice.[535] [536]

[522] Samuel Forman, Matthew Kelliher, and Gary Wood, "Clinical Improvement with Bottom-line Impact: Custom Care Planning for Patients with Acute and Chronic Illnesses in a Managed Care Setting," *The American Journal of Managed Care*, 3 (1997): 1041.

[523] Ingenix Corporation, "Identification and Management of High Risk Patients Using a Claims-based Predictive Model."

[524] John Lynch, Samuel A. Forman, Sandy Graff, and Mark C. Gunby, "High-risk Population Health Management – Achieving Improved Patient Outcomes and Near-term Financial Results," *The American Journal of Managed Care*, 6 (2000): 782.

[525] Goodwin, 65.

[526] Henry Dove, Ian Duncan, and Arthur Robb, "A Prediction Model for Targeting Low-cost, High-risk Members of Managed Care Organizations," *The American Journal of Managed Care*, 9 (2003): 385.

[527] Thomas Perls and Elizabeth R. Wood, "Acute Care Costs of the Oldest Old: They Cost Less, Their Case Intensity is Less, and They Go to Nonteaching Hospitals," *Archives of Internal Medicine*, 156 (1996): 754.

[528] Chap T. Le, Ping Liu, Bruce R. Lindgren, Kathleen A. Daly, and G. Scott Giebink, "Some Statistical Methods for Investigating the Date of Birth as a Disease Indicator," *Statistics in Medicine*, 22 (2003): 2128.

[529] Chap T. Le, 2127-8.

[530] Perls, 759.

[531] Perls, 759.

[532] See Xiaofeng Liu, Roland Sturm, and Brian J. Cuffel, "The Impact of Prior Authorization on Outpatient Utilization in Managed Behavioral Health Plans," *Medical Care Research and Review*, 57 (2000): 182.

[533] Implied, Liu, "The Impact of Prior Authorization on Outpatient Utilization in Managed Behavioral Health Plans," 185, 192.

[534] K. T. Palmer, M. J. Griffin, H. E. Syddall, A. Davis, B. Pannett, and D. Coggon, "Occupational Exposure to Noise and the Attributable Burden of Hearing Difficulties in Great Britain," *Occupational and Environmental Medicine*, 59 (2002): 634.

[535] Dimitri Christakis, Anne Kazak, Jeffrey Wright, Frederick Zimmerman, Alta Bassett, Frederick Connell, "What factors are associated with achieving high continuity of care?" *Family Medicine*, 36 (2004): 55-57.

- Unstructured text can be mined by a PM-like application to identify an individual at higher risk for using extended health services.[537]
- If unique values from records are removed, or entire claims records are deleted or obfuscated, PM may not identify such individuals. Disease management cost savings to the insurer might be undermined.

3.5.3.1 Security and Weights for Linear Regression Predictive Model

Such fields can be deidentified while preserving the above analyses. We explore the methods involved using one common PM approach, linear regression.[538] When PM is predicting costs or illness, the different data items analyzed are represented by numerical weights. The weights are combined to compute a total weight, which represents an individual's future predicted risk. PM can compare this value to a threshold and decide to enroll the individual in a disease management program. To deidentify the zip code field the entire 5-digit zip code could be encrypted and a lookup table be created, mapping the encrypted zip code to its associated weight. Zip codes would be identically encrypted in claim records during claim record deidentification before the claim records arrive at the copy data store. PM would obtain an encrypted zip code from a claim record and look up its ciphertext in the table to find the weight, which should be added to the total weight, which in turn can be compared to a threshold. Note, although HIPAA implies, via the reidentification code discussion in Section 3.1.1, that generating pseudonyms as a *function* of the original data should not be allowed we once again reference the argument in that section. If a statistician determines that there is a very low risk of reidentifying data subjects the transformation function should be permissible. We rely on such an argument here, i.e. using HIPAA's statistical method, as the mechanism for zip code deidentification and the other deidentifications which we will discuss below.

The same deidentification process as for patient zip code can be followed regarding a patient's name. Nanchahal creates directories describing the ethnic origin corresponding to a person's name.[539] Nanchahal studies how South Asian people's names may identify their South Asian origin. A regression weight can be assigned to each individual's name, representing the risk associated with her race.[540] The race weight could be obtained from other health services research, outside of Nanchahal's work, quantifying the risks of a race. To deidentify the lookup operation, the name in the claim records and in the directories, which could be constructed to map names to associated risk weights, can be deterministically encrypted with the same encryption function during claim record

[536] Jonathan P. Weiner, Stephen T. Parente, Deborah W. Garnick, Jinnet Fowles, Ann G. Lawthers, R. Heather Palmer, "Variation in Office-Based Quality: A Claims-Based Profile of Care Provided To Medicare Patients With Diabetes," *JAMA*, 273 (1995): 1503.

[537] Daniel Heinze, Mark L. Morsch, and John Holbrook, "Mining Free-text Medical Records" (Proceedings of the American Medical Informatics Association, 2001), 254-8.

[538] Ingenix Corporation, "Identification and Management of High Risk Patients Using a Claims-based Predictive Model."

[539] Nanchahal, 279.

[540] See Nanchahal, 279.

deidentification. Upon encountering an encrypted name in a claims record PM can look up the proper weight for the individual's race risk in a table, which can be added to the total weight.

The month and day of dates can also be deidentified. The length of activity can be provided in the UB92 instead of the "from" and "to" dates, as Enhancement 4 describes. This date difference would address the need to know the length of an activity. However, computing date order and date difference across multiple claim records can be important, too, as described in Section 3.5.3. The month and day of dates can be encrypted during claim record deidentification. The ciphertexts created, representing January 1 through December 31, can be placed behind an access control in a list in the order as the chronological order of the plaintexts. A function can be set up behind the access control. PM would take two encrypted dates and their unencrypted years from one or more claim records when processing the deidentified copy data store, and call the function behind the access control indicating whether it wants date order or date difference computed. If PM needs date order, the function would examine the supplied unencrypted years of the dates. If the years are different, the function would return to PM the date with the later year. It represents the later date. If the supplied years are the same, the function would compare the orders of both enciphered day-month values with the ordered list. It would return to PM the ciphertext of the date closest to the end of the list, which represents the later date. To obtain date difference, the function would subtract the order of one date from the order of the other using the orders in the enciphered chronological list. It would return to PM the difference, in total number of days, including an additional 365 days for each year difference between the unencrypted date years.

Other variables can be deidentified similarly. Date of birth (DOB) can be handled like the patient zip code field. The day, month, and year, or just year above 90, if more granularity is not necessary for PM, can be encrypted. A table can be set up mapping the encrypted DOBs to their associated risk weights. DOBs would be encrypted in claims records during claim record deidentification. PM can look up the encrypted DOBs it finds in the claims to find the needed weights, which would be added to the total weight. The presence of an authorization can be preserved by encrypting the authorization field.[541] The employer name field can be handled like the patient zip code field. It can be encrypted and a table be made available indicating the weight assigned to each employer. The employer name would be encrypted. PM could look up encrypted employer names in claims records in the created table while processing the deidentified claims data to find the needed weights. These weights could represent the employers' risks. Employer address can also be handled like patient zip code: parts of employer address (e.g., zip code) can be encrypted and a table mapping ciphertexts to weights can be created. PM would look up the encrypted parts in the claims data in the table to obtain the needed weights. Provider information, including the provider identifiers and the zip code in the Provider Name/Address/Phone Number field could also be handled like the above fields. A table can be made available mapping enciphered provider information to corresponding weights. Provider variables would be encrypted during claim record

[541] Liu, "The Impact of Prior Authorization on Outpatient Utilization in Managed Behavioral Health Plans," 187.

deidentification. PM would look up the weight when processing the deidentified claims and add the weight to the total weight. The logic for analyzing and creating the weight for free-form text can be installed earlier. The weight can be computed during claim record deidentification. During deidentification, the claim record can be analyzed and the computed weight provided in some location within the claim record. The free-form text data would be deleted. The provided weight can be added to the total weight later by PM. Unique values could be encrypted, and a lookup table mapping ciphertexts to weights can be created as for patient zip code. The values would be encrypted in claims records during claim record deidentification. PM would find the encrypted values in claim records when processing the copy data store and look them up to obtain the needed weights. Some of the deidentified items making a tuple unique in claims data can be similarly encrypted. A lookup table with needed weights can be created. The values making the tuple unique would be encrypted during claim record deidentification and PM would look up the needed weights when processing the deidentified claims.

The weight usage process described above can be made more secure. One security problem is that hiding plaintexts behind encryption yet exposing the needed weights to PM may allow for reidentification of the encrypted fields. For example, if somehow the weight associated with a zip code can be obtained from published literature, the zip code plaintext can be recoverable. One can simply look up the weight from the literature in the table available to PM and find the corresponding zip code pseudonym. By locating that pseudonym in the deidentified claims, the zip code value for those records has now been reidentified. The plaintext value from the literature underlies the pseudonym. To address this problem, one could place all the tables and all the weight mappings behind an access control. Instead of looking up each weight singularly, PM would work with a special function which would operate on a claims record in its entirety. The function would compute all the individual weights behind the access control and provide to PM a sum. Plaintext recovery has been considerably mitigated because a given weight is concealed by the sum, provided to PM, of several weights, and individual weights should no longer be discernable.

3.6 Chapter Conclusion

We have now shown that technically we can accomplish stronger privacy protection within applications like the BQMA. A linkage mechanism was introduced which securely improved the linkage of records in software applications. Also, we have demonstrated how a number of current installations of PM within insurance organizations relying on basic data, as well as PM platforms using more diversified data, may be run in a deidentified fashion. Since PM is operationally similar to the other BQMA, as shown in Section 1.3.2, the other BQMA may also be run in such a manner. In addition, other insurance applications might also be deidentified. Insurance organizations might have 12 or more different applications operating and computing like the BQMA. This includes the BQMA. Besides the BQMA, the insurer may have several other claims applications each with an independent data source: claims data capture, fraud analysis, claims adjudication, claims reports/document (such as the explanation of benefits sent to policy-holders), coordination of benefits, and premium setting.[542] [543] [544] [545] [546] Data sets containing enrollment data and policy-holder surveys may be additional data sets within an insurance organization with their own applications.[547] Since many of these applications process claims data and all of them need to link data, the BQMA and similar major applications within insurance organizations may be run in a privacy-preserving manner. Internal identifiable PHI can be better secured.

[542] Tricare, "Tricare Benefits for College Students," 26 February 2004, <http://www.tricare.osd.mil/collegestudents/TRICAREClaims.cfm> (5 September 2005).

[543] TMG Health, "Services," <http://www.tmghealth.com/services/claims.html> (5 September 2005).

[544] Interim Healthcare, "Retrospective Claims Analysis," <http://www.interimhealthcare.com/biz/CorpRetailPharma/EmpWellnessProgram/RetrospectiveClaimsAnalysis.asp> (5 September 2005).

[545] Tufts Health Plan, "Claims Procedures," *Billing Guidelines*, <http://www.tuftshealthplan.com/providers/provider.php?sec=administrative_resources&content=b_COB&rightnav=billing> (5 September 2005).

[546] Health Data Management, "Easy Access to Data Helps Insurers Make Timely Decisions," <http://www.healthdatamanagement.com/html/guide/toptech03alisting.cfm?AdvertiserID=753> (5 September 2005).

[547] America's Health Insurance Plans, "Personal Health Information, Health Plans, and Consumers."

4 Thesis Conclusion and Future Research

This thesis has demonstrated people's concern regarding medical privacy. At times, health organizations do not protect patient data. There are environmental, economic, organizational, and technical factors which appear to encourage less privacy protection. Legislation does not appear to encourage certain privacy protections. The economic costs and especially benefits of improved privacy protection are hard to quantify. Analysis of the impact of incorporating privacy protections on quality of care is unclear. Current technical approaches do not allow for easy linkage of erroneous data in a privacy-preserving way, an important utility for healthcare software applications. We showed how such beliefs may be reversed when privacy-related data are analyzed in more depth. We showed how pending 2005 legislation may be passed to encourage additional privacy protections; demonstrated how financial benefits of providing extra privacy enhancements for a set of key routine insurance software applications may exceed implementation costs for those applications within nine, but, most likely, considerably fewer years; explained how adopting a privacy-enhancing technology within those routine applications might improve care for policy-holders; and created techniques for protecting data while allowing for record linkage. A cryptographic threat model was created demonstrating how to evaluate security solutions when linking data obtained from internal or external organizations. More security must be provided if data are internally generated because the security protocol must guard against "internal" knowledge. We demonstrated how Predictive Modeling, one key insurance application among the routine applications we examined, could be run in a deidentified manner.

4.1 Future Work

Future research can expand upon our work. Our cost model can be elaborated. The valuation for deidentifying applications such as the BQMA to prevent the misuse of identifiable data could be expanded. The losses we used for computation only involved the costs to restore normal IT operations and manage the confidentiality breach.[548] [549] The losses only focused on attacks from internal employees. Losses were based on all employees in the organization regardless of whether they could, in fact, misuse data. A more robust valuation would incorporate all financial impacts properly apportioned. Total losses would hopefully include intangible losses. If data can only be abused by internal staff, total insider losses should only be divided by the number of employees who have access to identifiable data. If external attacks are possible, then external losses should obviously be included.

The number of data sets within the organization and the characteristics of the attackers could also enter the loss valuation. If the data within an organization are centralized in a single

[548] Computer Security Institute, "2005 CSI/FBI Computer Crime and Security Survey," 15.

[549] CRICO/RMF data.

data source, such as in a data warehouse, the valuation for securing the data would be the total losses suffered by the organization. All attacks against identifiable data should be mitigated as the single data source has been secured. Therefore, the valuation would be based on all potential losses. If there are multiple independent data sources, dividing the total number of losses by the number of data sources would yield a loss per every data set. Imagine an insurer suffers a loss of $250,000 because an employee obtained and publicized sensitive internal PHI.[550] To incorporate such a cost into the cost model, the $250,000 would have to be divided by 12 or more different data sources to compute the benefit of adding privacy protection to every data set. Recall 12 is roughly the number of data sets an average health insurer may have, as per Section 3.6. Protecting each data set would thus provide the corresponding *fraction* of the overall benefit to the organization. The valuation of privacy protection will be higher if the attackers possess characteristics leading to greater breach success. The attackers might have considerable financial resources, be technical experts, or be focused on a particular data set to the exclusion of others. A higher percentage of total losses could be apportioned to particularly vulnerable data or type of individuals as the probability of their attack success on particular data is greater.

4.1.1 Confirming Consumer Behavior

Social science research can verify the behavioral assumptions in our model. Surveys or experimental designs can be used. Will a small percent of individuals switch to the health plan of the insurer that provides additional BQMA privacy protection within a competitive insurance market? Will about 3.02% of individuals stop paying an average of $180 (2001 dollars) out of pocket and start submitting claims to an insurer if it provides extra BQMA privacy protection? To answer this question we must first confirm if the roughly $180 willingness-to-pay (WTP) behavior shown for 2001 also happens in 2005 or 2006. That is, is the behavior current? One concern with WTP behavior is poor survey response. People state their WTP based on what they may feel are "desired" responses, or cannot remember what they paid to protect their privacy when asked.[551] Given that salient privacy concerns continue today, including individuals' propensity to action as described in Section 2.2.1.1, current out-of-pocket payments should be like the California HealthCare Foundation Survey's 1999 results.

Can the insurer recover the individuals' WTP based on the accounting and health benefit redesigns it may undertake, as discussed in Section 2.2.3.1? This answer is also yes. First, in general, in business, past sales are strong predictors of future sales.[552] People should pay today what they paid for similar products before. Past payments for privacy

[550] For example, a hospital suffered a loss for this amount for the same reason: "Jury Orders Hospital to Pay $250,000 for Invasion of Privacy," *Aids Policy & Law*, 15 (2000): 10.

[551] Yaniv Poria and Harmen Oppewal, "Student Preferences for Room Attributes at University Halls of Residence: An Application of the Willingness to Pay Technique," *Tourism and Hospitality Research*, 4 (2002): 119.

[552] Tutor2u, "Sales Forecasting," <http://www.tutor2u.net/business/marketing/sales_forecasting.asp> (5 September 2005).

preservation may predict future privacy protection payments. Consider adolescents. When such individuals receive medical care, a variety of privacy concerns can arise, including those with billing or reimbursement procedures, scheduling notification, and privacy of medical records.[553] To alleviate such concerns, a youth can visit a variety of health care settings, including community centers, school-based and school-linked health clinics, and family planning clinics.[554] All these institutions can protect against the same privacy concerns. For example, the federal Title X Family Planning Program includes strong confidentiality protections at payment rates based on a sliding fee scale of the adolescent's (not his parents') income.[555] If he's uncomfortable at one Title X institution, the youth can visit a different one. He would pay the same amount to get the same privacy protection. Such behavior suggests that in a supportive context, paying one sum to protect privacy at one time may be indicative of an individual's capability in paying a similar sum to protect privacy at another time. Consequently, people may pay the insurer the same amount to protect their privacy as they used to pay their doctors out-of-pocket to protect the same kind of privacy.

In addition, in the insurance context, the insurer might need to recover the WTP to provide consistent health insurance. An additional roughly $4 million annual expense, as described before, may prevent the insurer from offering stable coverage to policy-holders due to this large expense. Since policy-holders want insurance, as they purchased it, the insurer can request they pay their WTP through one of the methods discussed before. This would be particularly true in a marketplace where it is the dominant insurer. The insurer might curtail coverage otherwise as it lacks the funding to provide consistent coverage.

How will women with preterm labor or individuals with other privacy-sensitive conditions behave? Will they pay out of pocket or avoid care at the beginning of their medical condition to maintain privacy protection or will they pay for the specific more "sensitive" medical visits during the care continuum to protect privacy just during those times?

4.1.2 Research to Frame "Partial" Privacy Protection

Critical to these questions is the notion of BQMA privacy protection. I refer to such privacy protection as *partial* privacy protection. BQMA privacy may be protected, but we do not know if other applications within the organization protect privacy. They might offer less privacy protection. Future research can examine the notion of incremental privacy provision. Organizations may not always protect privacy at all times due to costs, logistics, or unharmonized standards as this thesis has indicated. How do individuals perceive variable privacy protection? Do people believe organizations protect privacy based on the *percent* of the organization's internal software applications that protect

[553] Ford, 162.
[554] Ford, 165.
[555] Ford, 165.

privacy? If a certain minimum percent of applications is protected, for example 80%, is the organization seen as protecting privacy? Do people believe organizations protect privacy based on the *types* of applications that are protected? It's important for organizations to protect customer data for marketing or medical research applications. Protecting privacy within an actuarial application is less critical.

Similar to such a framing question, do people understand the effectiveness of given privacy protections? Many privacy definitions exist and many technical approaches protect data. Do consumers perceive the different risks based on an organization's privacy practices? Consider the Ethical Force Program (EFP) recommended privacy protections from before. What if, due to unharmonized standards or due to an organization's gradual implementation of privacy-protective practices, an organization at first only offers the Transparency aspect of all the EFP tenets, a notice of privacy practices? Would this be sufficient privacy protection for some individuals?[556] What if due to the same implementation issues the organization offers the EFP tenets of Transparency, Consent, Security, Data Quality, and Collection Limitation, but does not offer the tenets of Individual Access, Information Use Limitation, or Accountability, as described before? Is this appropriate privacy protection? Is privacy preserved only when all EFP tenets are enforced? Consider technical approaches. How does an organization explain to consumers that it is safeguarding privacy when the algorithms used are secure in certain data contexts but insecure in others? For example, consider the *ciphertext-only attack.* Imagine that an organization uses deterministic encryption to securely encrypt customers' last names. If clients have last names that are uniformly distributed among all the clients, the users of the deidentified data would not recognize the frequencies of particular last names because all frequencies should be similar. Deterministic encryption, which preserves plaintext distributions, is secure in this context because customer last names cannot be identified. If the organization acquires new clients, and the frequencies of these clients' last names differ—perhaps they are of a particular ethnic background and certain last names are much more prevalent than others—deterministic encryption may now allow the data users to guess the last names of some customers. The distributions of the enciphered last names would match the distributions of the original plaintexts because, again, deterministic encryption preserves the distributions. Would consumers understand such risks? What if a particular security protocol used to be secure, but was then "broken" because the cost of computational power declined making such a protocol *breakable*, or because someone discovered an actual flaw in the protocol? How is risk and security perceived by consumers?

4.1.3 Future Technical Enhancements

Future work can expand on a variety of technical issues. From the theoretical perspective on how to improve record linkage, the different distributions of the namespace of a field and the frequency of a field within a training file can be incorporated into our record linkage model. Weight and threshold computations would improve. For example, in our current

[556] See for instance Federal Trade Commission, "Online Profiling: A Report to Congress, Part 2 Recommendations," July 2000, <http://www.ftc.gov/os/2000/07/onlineprofiling.htm> (6 September 2005).

model the namespace is uniformly distributed. As a result, the probability of two fields in two records equaling when the records don't match is $(1/p_1)^{n1}$ where n1 is the number of character positions in the field and p_1 is the number of uniform values per position. A more realistic representation would note the frequency of different values for the specific field in a training file, e.g., $f_1,\ldots,f_k$. The possibility of two fields equaling in a validation file when their records don't match would then be approximately

$$\sum_{q=1}^{k} f_q^2$$

The first f_q is the probability of selecting that particular field value in the validation file and the second f_q is the probability of selecting another equal field value. The theoretical construction deciding whether to employ a string comparator for a field can be elaborated, going beyond the numerical example we provided. Recall one weight construction is:

$$\begin{aligned} W_{comparator} = \quad & \log (P (\text{convert } (\mathit{compare}(R_i[K], R_j[K])) = z \mid \\ & \qquad\qquad R_i \text{ and } R_j \text{ match}) / \\ & P (\text{convert } (\mathit{compare}(R_i[K], R_j[K])) = z \mid \\ & \qquad\qquad R_i \text{ and } R_j \text{ don't match})) \end{aligned}$$

Given a particular error and string comparator, a theoretical specification for $W_{comparator}$ can be provided just as for $[K_1]W_{agree}$ and $[K_1]W_{disagree}$ as before. These results can be extended with a theoretical or empirical investigation of $W_{comparator}$ when each field character is encrypted using our approach with padding. $W_{comparator}$ will differ. When computing $W_{comparator}$ for a field, the padding should come from a field different from the field being compared. Otherwise matching using character-level analysis would perform like full-field matching. Any two fields being compared would precisely not equal when one or both fields are in error regardless of whether the fields are encrypted at the character level or fully encrypted. A full field becomes the basis for an enciphered comparison because it is fully enciphered or fully used as padding. The error of the field used for padding and the original field's error would have to be simultaneously examined because the errors would interact. The software "middle layer" to provide the ability to compute results based on deidentified data yet restrict analysis to prevent discovery of additional information about units could be created. Basic primitives such as addition, equality, or order of deidentified data could be provided. These primitives can be invoked directly or a toolkit can be created.

From a cryptographic point of view, one interesting idea would be to create a secure hash or encryption function that preserves differences between underlying plaintexts. Such a technique would be more elegant than our approach because no padding is used. Akin to string comparators, a metric might be devised describing the similarity of ciphertexts based on a function describing similarity of plaintexts. Security would have to be redefined since encryption or hashing schemes typically transform slightly different plaintexts into fully different ciphertexts. Plaintext similarity is purposefully destroyed. In this case, some local plaintext structure must be preserved as some messages now have to be labeled as "close" to others.

4.2 Thesis Conclusion

As health insurance organizations increase the scope of their IT activities, privacy concerns may undermine decision-making. People's defensive behavior and data errors are creating suboptimal data. Technical and contextual factors are coalescing to encourage insurers to provide better privacy protection. The use of privacy enhancing technologies is becoming possible for applications. Such approaches should be considered in earnest by insurance managers. Otherwise, insurers might suffer financial, environmental, and quality of care consequences, as this thesis has shown.

This research, however, is useful beyond the health insurance marketplace. Other health care organizations may suffer similar consequences. Other health organizations would adopt technology for similar reasons as health insurers because in the health industry many organizations may have similar goals, as illustrated in Section 1.4. Other health organizations also use quality control applications like Utilization Review, Predictive Modeling, etc. This thesis has shown the benefits of adding stronger privacy protections to such applications: probable profitability, potentially less regulation, and enhanced patient health status. Other health organizations would want to adopt such protections in these and similar internal applications, too, to satisfy their aims. Our frameworks can be extended to other contexts. The secure linkage mechanism we created is a generic approach and is thus applicable to a wide variety of software applications. Testing robustness of Predictive Modeling for disease management can be extended to testing other data mining applications involving different methods of deidentification. Methodologies for testing the practicality of deidentified computation and analysis can improve. Frameworks quantifying the impact of data quality on high-risk pregnant women can be expanded to quantify the impact of data quality on e-commerce, national security, fraud analysis, and other information-driven domains.

In this thesis, we have constructed frameworks and technologies to improve organizational decision-making regarding privacy protection and information. We have shown how benefits provided by strategic applications to organizations are improved by improving underlying data quality. Our results show the enhanced business operations organizations can expect by improving consumer privacy. Privacy protection offers unmistakable value to organizations.

Glossary

AAE = Attacks Against Encryption. A set of realistic cryptographic attacks against software applications protected by encryption within an organizational context based on theoretical encryption notions.

BFA = Brute Force Attack. Traversing through all original messages or all cryptographic keys with regard to an encryption or hashing scheme to compromise the scheme by finding the encryption or hashing key or some or all of the original plaintext messages.

BQMA = Basic Quality Management Applications. A set of software applications used by health insurance organizations to process patient data for organizational quality control purposes.

EFP = Ethical Force Program. A policy group related to the American Medical Association that created a list of managerial, operational, and technical principles describing how to protect health information within organizations.

HEDIS = Health Plan Employer Data and Information Set. A health insurance organization's operational data used to quantify the insurer's performance regarding its coverage and services.

HHS = US Department of Health and Human Services.

HIPAA = Health Insurance Portability and Accountability Act. The federal law describing the protection that must be offered to health information by "covered" health organizations, such as health insurers and providers, which manage this information.

NICU = Neonatal Intensive Care Unit. A specially equipped nursery to care for ill newborns.

OCR = Office for Civil Rights. US Department of Health and Human Services office that enforces the Health Insurance Portability and Accountability Act.

PHI = Protected Health Information. Any individually-identifiable health information relating to the data subject's medical care, the delivery of that care, or the payment for that care (as defined by the Health Insurance Portability and Accountability Act).

PM = Predictive Modeling. One of the software platforms making up the Basic Quality Management Applications. It identifies individuals for disease management, which is a set of programs attempting to care for chronically ill or "high-risk" individuals and reduce the costs of their care.

UB92 = A type of health care claims record used by health insurance companies. Such records may be used by PM or other BQMA software.

WTP = Willingness To Pay. A social science term. An individual's stated, as opposed to revealed, payment preference for a particular item or service.

Bibliography

1) Center for Democracy and Technology. "Statement of Janlori Goldman, House Committee on Government Reform and Oversight." 1996. <http://www.cdt.org/testimony/960614goldman.html>.

2) Health Privacy Project. "Health Privacy Polling Data." 2001. <http://www.healthprivacy.org/content2310/content.htm>.

3) Goldman, Janlori and Zoe Hudson. "Virtually Exposed: Privacy and E-health." *Health Affairs*, 19 (2000): 140-8.

4) Harris Interactive. "Privacy On and Off the Internet: What Consumers Want." 2002. 1-127, <http://www.aicpa.org/download/webtrust/priv_rpt_21mar02.pdf>.

5) HIPAAps Privacy and Security. "Examples of Privacy Violations." 2003. <http://www.hipaaps.com/main/examples.html>.

6) Hoffman, Donna L. "The Consumer Experience: A Research Agenda Going Forward." 14 May 2003. <http://elab.vanderbilt.edu/research/papers/pdf/manuscripts/FTC.privacy.pdf>.

7) Dinev, Tamara. "Privacy Concerns and Internet Use – A Model of Tradeoff Factors." <http://wise.fau.edu/~tdinev/publications/privacy.pdf>.

8) Caudill, Eve M and Patrick Murphy. "Consumer Online Privacy: Legal and Ethical Issues." *Journal of Public Policy & Marketing*, 19 (2000): 7-19.

9) Secretariat, Treasury Board of Canada. "So, What Exactly Is Privacy?" 26 September 2003. <http://www.cio-dpi.gc.ca/pgol-pged/piatp-pfefvp/course1/mod1/mod1-2_e.asp>.

10) Smith, H. Jeff, Sandra J. Milberg, and Sandra J. Burke. "Information Privacy: Measuring Individuals' Concerns About Organizational Practices." *MIS Quarterly*, 20 (1996): 167-196.

11) Centers for Disease Control and Prevention. "HIPAA Privacy Rule and Public Health." 2003. <http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm>.

12) US Department of Health and Human Services. "Standards for Privacy of Individually Identifiable Health Information" (part 1). 28 December 2000. 82462-82510, <http://www.hhs.gov/ocr/part1.pdf>.

13) Health Privacy Project. "Exposed: A Health Privacy Primer for Consumers." 1999. 1-16, <http://www.healthprivacy.org/usr_doc/34775.pdf>.

14) Health Privacy Project. *State Privacy Law Summaries*. <http://www.healthprivacy.org/info-url_nocat2304/info-url_nocat_search.htm>.

15) American Association of Health Plans. "About AAHP." 2003. <http://www.aahp.org/template.cfm?section=About_AAHP>.

16) American Association of Health Plans. "Statement on the Confidentiality of Medical Information and the Medical Information Protection Act of 1998." 1998. <http://www.aahp.org/Content/ContentGroups/Testimony/Confidentiality_and_Protection_of_Medical_Information_(Feb__26,_1998).htm>.

17) National Association of Insurance Commissioners. "NAIC Mission Statement." <http://www.naic.org/about/mission.htm>.

18) American Medical Association. "The Ethical Force Program." December 2000. 1-30, <http://www.ama-assn.org/ama/upload/mm/369/ef_privacy_rpt.pdf>.

19) National Research Council. *For the Record: Protecting Electronic Health Information*. Washington, DC: National Academy Press, 1997.

20) Goldman, Janlori. "Protecting Privacy to Improve Health Care." *Health Affairs*, 17 (1998): 47-60.

21) US Department of Health and Human Services. "Standards for Privacy of Individually Identifiable Health Information" (update). 14 August 2002. 53182-53273, <http://www.hhs.gov/ocr/hipaa/privrulepd.pdf>.

22) American Medical Association. "About the Ethical Force Program." 18 July 2005. <http://www.ama-assn.org/ama/pub/category/14401.html>.

23) Wynia, Matthew K., Steven S. Coughlin, Sheri Alpert, Deborah S. Cummins, Linda L. Emanuel. "Shared Expectations for Protection of Identifiable Health Care Information." *Journal of General Internal Medicine*, 16 (2001): 100-11.

24) The Kansas Department of Health and Environment. "Charitable Health Program Overview." <http://www.kdhe.state.ks.us/olrh/CHPoverview.htm>.

25) W.K. Kellogg Foundation. "Frequently Asked Questions…Insurance and Managed Care." 1-15, <http://www.wkkf.org/Pubs/Devolution/NCSL_FA_Insurance_and_managed_care_00331_02768.pdf>.

26) American Health Information Management Association. "Sizing up HEDIS: Experts Take System's Measure." 2002. <http://library.ahima.org/xpedio/groups/public/documents/ahima/pub_bok1_009714.html>.

27) FACTS Services, Inc. "Products At-a-glance." <http://factsservices.com/products/products_glance.asp>.

28) DxCG. "Disease Management and Quality Improvement Report." May 2003. <http://www.dxcg.com/press/DMQualityReport.pdf>.

29) Ingenix Corporation. "Improve Your Medical Management and Underwriting Effectiveness." <http://www.ingenix.com/esg/products.php?pid=10>.

30) Enthoven, Alain C and Sara Singer. "The Managed Care Backlash and the Task Force in California." *Health Affairs*, 17 (1998): 95-6.

31) Kremer, Thomas G. and Ellis Gesten. "Confidentiality Limits of Managed Care and Clients' Willingness to Self-Disclose." *Professional Psychology: Research and Practice*, 29 (1998): 553-8.

32) Bettermanagement.Com. "Effective Provider Profiling: Enhancing Care, Improving Costs." webcast.

33) National Committee for Quality Assurance. "HEDIS 2005 Summary Table of Measures and Product Lines." <http://www.ncqa.org/Programs/HEDIS/HEDIS%202005%20Summary.pdf>.

34) National Committee for Quality Assurance. "The Health Plan Employer Data and Information Set (HEDIS)." <http://www.ncqa.org/Programs/HEDIS/>.

35) Thompson, Joseph W., Sathiska D. Pinidiya, Kevin W. Ryan, Elizabeth D. McKinley, Shannon Alston, James E. Bost, Jessica Briefer French, and Pippa Simpson. "Health Plan Quality-of-Care Information Is Undermined by Voluntary Reporting." *American Journal of Preventive Medicine*, 24 (2003): 62-70.

36) Disease Management Association of America. "Definition of Disease Management." 2003. <http://www.dmaa.org/definition.html>.

37) Case Western Reserve University. "Disease Management Programs." <http://www.case.edu/med/epidbio/mphp439/Disease_Management.htm>.

38) Privacy Sector Advocacy. "Disease Management and Chronic Diseases." November 2002. 1-5.

39) Norris, Susan L., Phyllis J. Nichols, Carl J. Caspersen, Russell E. Glasgow, Michael M. Engelgau, Leonard Jack, Jr, George Isham, Susan R. Snyder, Vilma G. Carande-Kulis, Sanford Garfield, Peter Briss, and David McCulloch. "The Effectiveness of Disease and Case Management for People with Diabetes. A Systematic Review." *American Journal of Preventative Medicine*, 22 (2002), 15-38.

40) California Healthcare Foundation. "E-disease Management." November 2001. 1-48, <http://www.chcf.org/documents/ihealth/EDiseaseManagement.pdf>.

41) Institute of Medicine. *Crossing the Quality Chasm*. Washington, DC: National Academy Press, 2001.

42) Government of British Columbia. "Chronic Disease and Your Health: Information for Patients," *Chronic Disease Management*. 2003. <http://www.healthservices.gov.bc.ca/cdm/patients/index.html>.

43) American Association of Health Plans/Health Insurance Association of America. "The Cost Savings of Disease Management Programs: Report on a Study of Health Plans." November 2003. <http://www.aahp.org/Content/ContentGroups/Homepage_News/Disease Management_Short_Report.doc>.

44) Pacific Business Group on Health. "Disease Management Effectiveness Project." November 2002. <http://www.pbgh.org/programs/dmep/disease_mgmt_report_11-02.pdf>.

45) Baker, Geoffrey B. "Integrating Technology and Disease Management: The Challenges." *Healthplan*, 43 (2002): 60-2, 64-6.

46) "Predictive Modeling, Integrated Disease Management Emerge as Popular Strategies." *Data Strategies and Benchmarks*, 6 (2002).

47) Welch, W. Pete, Christopher Bergsten, Charles Cutler, Carmella Bocchino, and Richard I. Smith. "Disease Management Practices of Health Plans." *The American Journal of Managed Care*, 8 (2002): 353-61.

48) Dalrymple, A. J., L. S. Lahti, L. J. Hutchison, and J. J. O'Doherty. "Record Linkage in a Regional Mental Health Planning Study: Accuracy of Unique Identifiers, Reliability of Sociodemographics, and Estimating Identifier Error." *Journal of Mental Health Administration*, 21 (1994): 185-92.

49) Capitation Management Report. "Are You Ready to Take on Claims Adjudication." September 1999. <http://www.phoenixservice.net/Articles/article6.pdf>.

50) HealthcareIndustryPulse. "Payment Errors Cost MCOs Big Money." January 2005. <http://www.bdo.com/about/publications/industry/hcp_jan_05/claims.asp>.

51) Centers for Medicare and Medicaid services. "Health Insurance Claim Form." <http://cms.hhs.gov/providers/edi/cms1500.pdf>.

52) Centers for Medicare and Medicaid Services. "Uniform Bill." <http://cms.hhs.gov/providers/edi/h1450.pdf>.

53) Government Accounting Office. "Medical Records Privacy: Access Needed for Health Research, But Oversight of Privacy Protections Limited." February 1999. <http://www.gao.gov/archive/1999/he99055.pdf>.

54) America's Health Insurance Plans. "Personal Health Plan Information, Health Plans, and Consumers." *AHIP Center for Policy and Research*. August 2001. <http://www.ahipresearch.org/PDFs/24_CRAfinalreportPriv-Conf.pdf>.

55) Privacy Rights Clearinghouse. "A Chronology of Data Breaches Reported Since the ChoicePoint Incident." 30 August 2005. <http://www.privacyrights.org/ar/ChronDataBreaches.htm>.

56) Radcliff, Deborah. "Invisible Loot." *Industry Week*, 2 November 1998, <http://www.industryweek.com/CurrentArticles/ASP/articles.asp?ArticleId=298>.

57) Katz, David. "Elements of a Comprehensive Security Solution." *Health Management Technology*, 21 (2000): 12-6.

58) US Department of Health and Human Services. "Standards for Privacy of Individually Identifiable Health Information" (part 2). 28 December 2000. 82511-82560, <http://www.hhs.gov/ocr/part2.pdf>

59) Iezzoni, Lisa I. *Risk Adjustment for Measuring Healthcare Outcomes, Second Edition*. Chicago, IL: Health Administration Press, 1997.

60) Healthcare Informatics Online. "Will Your Patient Data Merge With You?" 1997. <http://www.healthcare-informatics.com/issues/1997/04_97/merge.htm>.

61) Fernandes, Lorraine, Celia Lenson, Joe Hewitt, Jerry Weber, and Jo Ann Yamamoto. "Medical Record Number Errors." White Paper from Initiate Corporation, April 2001, 1-12.

62) Healthcare Information and Management Systems Society. "U.S. Healthcare Industry: HIPAA Compliance Survey Results: Winter 2005." <http://www.himss.org/Content/files/WinterSurvey2005.pdf>.

63) Carpenter, Vivian and Ehsan H. Feroz. "Institutional Theory and Accounting Rule Choice: An Analysis of Four US State Governments' Decisions to Adopt Generally Accepted Accounting Principles." *Accounting, Organizations, and Society*, 26 (2001): 565-596.

64) Hu, Paul Jen-Hwa, Patrick Chau, and Olivia Liu Sheng. "Investigation of Factors Affecting Healthcare Organization's Adoption of Telemedicine Technology" (Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000).

65) Rahim, Md. Mahbubur, G. Shanks, and R.B. Johnston. "Understanding Motivations for IOS Adoption" (Proceedings of the Twelfth Australasian Conference on Information Systems).

66) Zutshi, A. and A. Sohal. "Environmental Management System Adoption by Australasian Organizations: Part 1: Reasons, Benefits and Impediments." *Technovation*, 24 (2000): 335-357.

67) Rand Corporation. "How MCO Medical Directors See the System." *Managed Care and the Evaluation and Adoption of Emerging Medical Technologies*. 2000. 33, <http://www.rand.org/publications/MR/MR1195/MR1195.chap4.pdf>.

68) Rand Corporation. "How Might Technology Adoption be Improved." *Managed Care and the Evaluation and Adoption of Emerging Medical Technologies*. 2000. 49, <http://www.rand.org/publications/MR/MR1195/MR1195.chap6.pdf>.

69) Elliot, S.R. "Adoption and Implementation of IT: An Evaluation of the Applicability of Western Strategic Models to Chinese Firms." In *Diffusion and Adoption of Information Technology*, edited by Karlheinz Kautz. London: Chapman & Hall, 1996.

70) General Accounting Office. "Health Insurance Regulation: Varying State Requirements Affect Cost of Insurance." 1999. <http://www.gao.gov/archive/1996/he96161.pdf>.

71) Strum, Roland and J. Unutzer. "State Legislation and the Use of Complementary and Alternative Medicine." *Inquiry – Blue Cross and Blue Shield Association*, Winter 2000/2001, 423-9.

72) Managed Care Magazine. "State Mandates Promote Contraceptive Coverage." 2004. <http://www.managedcaremag.com/archives/0408/0408.formularyfiles.html>.

73) US Department of Health and Human Services. "Protecting the Privacy of Patients' Health Information: Summary of the Final Regulation." 2000. <http://www.hhs.gov/news/press/2000pres/00fsprivacy.html>.

74) University of Miami. "Privacy/Data Protection Project." 15 August 2002. <http://privacy.med.miami.edu/glossary/xd_consent.htm>.

75) American Academy of Ophthalmic Executives. "Final HIPAA Privacy Rule." <http://www.aao.org/aaoesite/promo/compliance/hipaa_final.cfm>.

76) Atlantic Information Services. "HIPAA Compliance Strategies." 2003. <http://www.aishealth.com/Compliance/Hipaa/MCWDMTraining.html>.

77) US Department of Health and Human Services. "Standards for Privacy of Individually Identifiable Health Information" (part 3). 28 December 2000. 82561-82610, <http://www.hhs.gov/ocr/part3.pdf>.

78) Alcohol, Drug, and Mental Health Board for Franklin County. "Info for Consumers." <http://www.adamhfranklin.org/consumers/hipaaPolicy05.php>.

79) Rovner, Jack A. "Don't Let Fear of HIPAA Keep You from Crucial Data." *Managed Care Magazine*. March 2003. <http://www.managedcaremag.com/archives/0303/0303.legal.html>.

80) Thompson, Mark S. *Decision Analysis for Program Evaluation*. Cambridge, MA: Ballinger Publishing Company, 1982.

81) Barr, Judith and Gerald Schumacher. "Using Decision Analysis to Conduct Pharmacoeconomic Studies." In *Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition,* edited by B. Spilker. Philadelphia: Lippincott-raven Publishers, 1996.

82) Hunink, Myriam and Paul Galsziou. *Decision Making in Health and Medicine: Integrating Evidence and Values.* Cambridge, UK: Cambridge University Press, 2001.

83) Wiley Rein & Fielding LLP. "A New Era for HIPAA Enforcement." May 2004. <http://www.wrf.com/publication_newsletters.cfm?sp=newsletter&year=2004&ID=10&publication_id=9825&keyword>.

84) iHealthbeat. "Enforcement of HIPAA Privacy: Making it Real." 19 November 2003. <http://ihealthbeat.org/index.cfm?Action=dspItem&itemID=100262>.

85) Smith, G. Stevenson. "Recognizing and Preparing Loss Estimates from Cyber-Attacks." *Information Systems Security,* 12 (2004); 46-57.

86) Soo Hoo, Kevin J. "How Much is Enough? A Risk-management Approach to Computer Security." June 2000, 40-41, <http://iis-db.stanford.edu/pubs/11900/soohoo.pdf>.

87) Health Privacy Project. "Best Principles for Health Privacy." July 1999. <http://www.healthprivacy.org/usr_doc/33807.pdf>.

88) Liu, Chang, Jack T. Marchewka, June Lu, and Chun-Sheng Yu. "Beyond Concern: A Privacy-Trust-Behavioral Intention Model of Electronic Commerce." *Information & Management*, 42 (2004):127-142.

89) Kenny, Steve and John Borking. "The Value of Privacy Engineering." *Journal of Information, Law, and Technology*, 1 (2002), <http://www2.warwick.ac.uk/fac/soc/law/elj/jilt/2002_1/kenny/>.

90) Wirtz, Bernd W. and Nikolai Lihotzky. "Customer Retention Management in the B2C Electronic Business." *Long Range Planning*, 36 (2003): 517-32.

91) Computer Security Institute. "2004 CSI/FBI Computer Crime and Security Survey." 2004. <http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2004.pdf>.

92) Computer Security Institute. "2001 CSI/FBI Computer Crime and Security Survey." Spring 2001. <http://www.reddshell.com/docs/csi_fbi_2001.pdf>.

93) Computer Security Institute. "2003 CSI/FBI Computer Crime and Security Survey." 2003. <http://www.reddshell.com/docs/csi_fbi_2003.pdf>.

94) Shostak, Adam. "'People Won't Pay for Privacy,' Reconsidered." 14 March 2003, <http://www.cpppe.umd.edu/rhsmith3/papers/Final_session3_shostack_privacy.pdf>.

95) Savage, Marcia. "Cybersecurity Boosts Bottom Line." *SC Magazine*, 13 February 2005, <http://www.scmagazine.com/news/index.cfm?fuseaction=newsDetails&newsUID=87605d0f-ffc6-4169-93e4-3c7274412de7&newsType=Latest%20News>.

96) Love, James. "NYT on OMB's Costs Benefit of Losses of Liberty and Privacy." 2 April 2003, <http://lists.essential.org/pipermail/random-bits/2003-April/001054.html>.

97) Tracy, C. Shawn, Guilherme Coelho Dantas, and Ross EG Upshur. "Feasibility of a Patient Decision Aid Regarding Disclosure of Personal Health Information: Qualitative Evaluation of the Health Care Information Directive." *BMC Medical Informatics and Decision Making*, 4 (2004), <http://www.pubmedcentral.gov/picrender.fcgi?artid=518970&blobtype=pdf>.

98) Stewart, Kathy A. and Albert H. Segars. "An Empirical Examination of the Concern for Information Privacy Instrument." *Information Systems Research*, 13 (2002): 36-49.

99) The Pew Internet & American Life Project. "Trust and Privacy Online: Why Americans Want to Rewrite the Rules." 20 August 2000, 1-29, <http://www.pewinternet.org/pdfs/PIP_Trust_Privacy_Report.pdf>.

100) Federal Committee on Statistical Methodology Confidentiality and Data Access Committee. "Restricted Access Procedures." 4 April 2002, 1-16, <http://www.fcsm.gov/committees/cdac/cdacra9.pdf>.

101) Siegel, Carol A., Ty Sagalow, and Paul Serritella. "Cyber-risk Management: Technical and Insurance Controls for Enterprise-level Security." *Information Systems Security*, 11 (2002): 33-50.

102) Rindfleisch, Thomas C. "Privacy, Information Technology, and Health Care." *Communications of the ACM*, 40 (1997): 92-100.

103) Friedman, Leonard and D.B. White. "What Is Quality, Who Wants It, and Why?" *Managed Care Quarterly*, 7 (1999): 40-6.

104) Joint Commission on Accreditation of Healthcare Organizations. "What is the Joint Commission on Accreditation of Healthcare Organizations?" 2005. <http://www.jcaho.org/general+public/who+jc/index.htm>.

105) Agency for Healthcare Research and Quality. "What is AHRQ?" February 2002. <http://www.ahrq.gov/about/whatis.htm>.

106) Sandrick, Karen. "Tops in Quality." *Trustee*, 56 (2003): 12-6.

107) President's Advisory Commission on Consumer Protection and Quality in the Health Care Industry. <http://www.hcqualitycommission.gov/>.

108) Born, Patricia H. and Carol Simon. "Patients and Profits: The Relationship between HMO Financial Performance and Quality of Care." *Health Affairs*, 20 (2001): 167-74.

109) McCullagh, Declan. "Congress Edges toward New Privacy Rules." *News.com*, 10 March 2005, <http://news.com.com/Congress+edges+toward+new+privacy+rules/2100-1028_3-5609324.html?tag=nl>.

110) Center for Democracy and Technology. "Consumer Privacy Legislation (109th)." <http://www.cdt.org/legislation/0/3/>.

111) The Henry J. Kaiser Family Foundation. "Health Insurance Coverage in America, 2001 Data Update." January 2003. <http://www.kff.org/uninsured/loader.cfm?url=/commonspot/security/getfile.cfm&PageID=14309>.

112) The Henry J. Kaiser Family Foundation. "Employer Health Benefits, 2001 Annual Survey." 2001. 1-150, <http://www.kff.org/insurance/loader.cfm?url=/commonspot/security/getfile.cfm&PageID=13836>.

113) AcademyHealth. "Mapping State Health Insurance Markets, 2001: Structure and Change." September 2003. <http://www.statecoverage.net/pdf/mapping2001.pdf>.

114) Chatzky, Jean Sherman. "Choosing Health Care." *Money*, 27 (1998): 152-60.

115) Culnan, Mary J. and George Milne. "The Culnan-Milne Survey on Consumer & Online Privacy Notices." December 2001. <http://www.ftc.gov/bcp/workshops/glb/supporting/culnan-milne.pdf>.

116) Harris Interactive. "Privacy Notices Research: Final Results." December 2001. <http://www.ftc.gov/bcp/workshops/glb/supporting/harris%20results.pdf>.

117) Allstate, Workplace Division. "Workplace Marketing." <http://www.ahlcorp.com/ProdIndWork.asp>.

118) "Living Large." *Health Management Technology*, 24 (2003): 32-33.

119) Benko, Laura. "Long-range Forecast: Partly Healthy, Chance of Storms." *Modern Healthcare*, 34 (2004): 26-30.

120) Atkinson, William. "Making Disease Management Work." *Society for Human Resource Management*, 47 (2002), <http://www.shrm.org/hrmagazine/articles/0102/0102atkinson.asp>.

121) Peel, Deborah. “Lawsuit Challenges HIPAA.” *MSPP News*, 13 (2003), <http://www.mspp.net/hipaa_lawsuit.htm>.

122) Federal Reserve Bank of Minneapolis. “What is a Dollar Worth?” <http://minneapolisfed.org/research/data/us/calc/>.

123) Unpublished data from Executive Information System database, Controlled Risk Insurance Company (CRICO)/Risk Management Foundation (RMF). Obtained on March 4, 2005.

124) Computer Security Institute. “2005 CSI/FBI Computer Crime and Security Survey.” <http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2005.pdf>.

125) CSO Magazine. “2004 E-Crime Watch Survey.” <http://www.cert.org/archive/pdf/2004eCrimeWatchSummary.pdf>.

126) “Fortune 500 Largest US Corporations.” *Fortune*, F-31, April 15, 2002.

127) Kirkman-Liff, Bradford. “Restoring Trust to Managed Care, Part 1: A Focus on Patients.” <http://www.ajmc.com/files/articlefiles/AJMC2003feb1Kirkman174-180.pdf>.

128) California Healthcare Foundation. “Medical Privacy and Confidentiality Survey.” 1999. <http://www.chcf.org/documents/ihealth/topline.pdf>.

129) Benko, Laura. “Less is Not More.” *Modern Healthcare*, 30 (2000): 40-4.

130) Benko, Laura. “…You Pay.” *Modern Healthcare*, 33 (2003): 8.

131) Grimaldi, Paul. “The Versatile Medical Loss Ratio.” *Nursing Management*, 29 (1998): 12-6.

132) Niedzielski, Joe. “Rising Expenses Nip at Results of Public HMOs.” *National Underwriter*, 100 (1996): 4, 9.

133) Harvard Pilgrim Health Care. “Annual Report 2004.” <http://www.harvardpilgrim.org/hpimages/HP-2004Annual.pdf?SMSESSION=NO>.

134) Blue Cross and Blue Shield of Massachusetts. “Annual Report 2004.” <http://www.bcbsma.com/common/en_US/repositories/CommonMainContent/aboutUs/AnnualReport/BCBSMA_04_Financials.pdf>.

135) Benko, Laura. “Earnings at a Premium.” *Modern Healthcare*, 32 (2002): 22-23.

136) Grimaldi, Paul. “Medical Loss Ratios under Scrutiny.” *Nursing Management*, 27 (1996): 14-7.

137) 6, Perry. “Who Wants Privacy Protection and What Do They Want?” *Journal of Consumer Behavior*, 2 (2002): 80-100.

138) Parente, Stephen, Roger Feldman, and Jon B. Christianson. “Employee Choice of Consumer-driven Health Insurance in a Multiplan, Multiproduct Setting.” *Health Services Research*, 39 (2004): 1091-1111.

139) Robinson, James. “Reinvention of Health Insurance in the Consumer Era.” *JAMA*, 291 (2004): 1880-6.

140) Dye, Timothy, Martha Wojtowycz, Mary Applegate, and Richard Aubry. “Women’s Willingness to Share Information and Participation in Prenatal Care Systems.” *American Journal of Epidemiology*, 156 (2002): 286-91.

141) Morrison, John, Niki K. Bergauer, Debbie Jacques, Suzanne K. Coleman, and Gary J. Stanziano. “Telemedicine: Cost-Effective Management of High-Risk Pregnancy.” *Managed Care*, 10 (2001): 42-49.

142) Corwin, Michael, Susan M. Mou, Shirazali G. Sunderji, Stanley Gall, Helen How, Vinu Patel, and Mark Gray. “Obstetrics: Multicenter Randomized Clinical Trial of Home Uterine Activity Monitoring: Pregnancy Outcomes for All Women Randomized.” *American Journal of Obstetrics and Gynecology*, 175 (1996): 1281-5.

143) Ross, Michael, Catherine A. Downey, Rose Bemis-Heys, Men Nguyen, Debbie L. Jacques, and Gary Stanziano. “Prediction by Maternal Risk Factors of Neonatal Intensive Care Admissions: Evaluation of >59,000 Women in National Managed Care Programs.” *American Journal of Obstetrics and Gynecology*, 181 (1999): 835-42.

144) Lear, Deanna, Laura C. Schall, Gary M. Marsh, Ken S. Liu, and Yvonne Yao. “Identification and Case Management in an HMO of Patients at Risk of Preterm Labor.” *The American Journal of Managed Care*, 4 (1998): 865-71.

145) Hutti, Marianne and Wayne M. Usui. “Nursing Telephonic Case Management and Pregnancy Outcomes of Mothers and Infants.” *Lippincott’s Case Management*, 9 (2004): 287-299.

146) Fangman, John J., Peter M. Mark, Leslie Pratt, Kathleen K. Conway, Margaret L. Healey, John W. Oswald, Donald L. Uden. “Prematurity Prevention Programs: An Analysis of Successes and Failures.” *American Journal of Obstetrics and Gynecology*, 170 (1994): 744-50.

147) Kempe, Allison, Benjamin P. Sachs, Hope Ricciotti, Arthur M. Sobol, and Paul H. Wise. “Home Uterine Activity Monitoring in the Prevention of the Very Low Birth Weight.” *Public Health Reports*, 112 (1997): 433-9.

148) MEDecision. “Advanced Medical Management” <http://www.medecision.com/page.cfm?page=advanced>.

149) Health Alliance Plan. “Healthy Living, Prenatal Care Chart.” <http://www.hap.org/healthy_living/teenadult/prenatca.php#High%20Risk>.

150) Koroukian, Siran and Alfred A. Rimm. “The ‘Adequacy of Prenatal Care Utilization’ (APNCU) Index to Study Low Birth Weight: Is the Index Biased?” *Journal of Clinical Epidemiology*, 55 (2002): 296-305.

151) Virtual Hospital. “Obstetrics: Prenatal Care.” *University of Iowa Family Practice Handbook, Fourth Edition.* <http://www.vh.org/adult/provider/familymedicine/FPHandbook/Chapter14/02-14.html>.

152) Sanin-Blair, J., M. Palacio, J. Delgado, F. Figueras, O. Coll, L. Cabero, V. Cararach, and E. Gratacos. “Impact of Ultrasound Cervical Length Assessment on Duration of Hospital Stay in the Clinical Management of Threatened Preterm Labor.” *Ultrasound in Obstetrics & Gynecology*, 24 (2004): 756-60.

153) King, Derek and Elias Mossialos. “The Determinants of Private Medical Insurance Prevalence in England, 1997-2000.” *Health Services Research*, 40 (2005): 195-212.

154) Janlori Goldman, Director, Health Privacy Project, <http://www.healthprivacy.org/info-url_nocat2301/info-url_nocat_show.htm?doc_id=33777>.

155) Health Privacy Project. “Testimony of Janlori Goldman before Senate Special Committee on Aging.” 23 September 2003. <http://www.healthprivacy.org/usr_doc/sentestimony.pdf>.

156) Kaisernetwork.org. “Daily Health Policy Report.” 2001. <http://www.kaisernetwork.org/daily_reports/rep_index.cfm?DR_ID=4609>.

157) Green, Jan. “Sizing Up the Sickest.” *Hospitals & Health Networks*, 72 (1998): 28-31.

158) Sloan, Mary Anne. “Targeting Populations at Highest Risk: IT Delivers a 2:1 ROI for Midwest Health Plan – What Works: Disease Management.” *Health Management Technology*. September 2003. <http://www.findarticles.com/p/articles/mi_m0DUD/is_9_24/ai_108148076>.

159) Muender, Melissa, Mary Lou Moore, Guoqing John Chen, and Mary Ann Sevick. “Cost-benefit of a Nursing Telephone Intervention to Reduce Preterm and Low-birthweight Births in an African American Clinic Population.” *Preventive Medicine*, 30 (2000): 271-6.

160) Ofman, Joshua J., Seonyoung Ryu, Jeff Borenstein, Stephen Kania, Jay Lee, Amy Grogg, Christina Farup, and Scott Weingarten. “Identifying Patients with Gastroesophageal Reflux Disease in a Managed Care Organization.” *American Journal of Health-system Pharmacy*, 58 (2001): 1607-13.

161) Centers for Disease Control and Prevention. “National Vital Statistics Report: Births: Final Data for 2002.” 1-116, <http://www.cdc.gov/nchs/data/nvsr/nvsr52/nvsr52_10.pdf>.

162) March of Dimes. “Multiples: Twins, Triplets, and Beyond,” *Professionals and Researchers*. <http://www.marchofdimes.com/professionals/14332_4545.asp>.

163) General Accounting Office. “Social Security: Government and Commercial Use of the Social Security Number is Widespread.” February 1999. 1-21, <http://www.gao.gov/archive/1999/he99028.pdf>.

164) General Accounting Office. “Social Security Numbers.” January 2004. 1-22, <http://www.gao.gov/new.items/d04768t.pdf>.

165) Winkler, William E. “Preprocessing of Lists and String Comparison.” In *Record Linkage Techniques* edited by W. Alvey and B. Kilss. U.S. Internal Revenue Service, 181, 1985.

166) American Civil Liberties Union. “ACLU Says CDC Guidelines on HIV Surveillance Could Lead to Better Privacy Protections.” *News*, 1999. <http://www.aclu.org/Privacy/Privacy.cfm?ID=8791&c=27>.

167) State of Florida, Agency for Health Care Administration. “The Florida Medicaid Disease Management Experience.” 26 January 2005. <http://www.fdhc.state.fl.us/Medicaid/deputy_secretary/recent_presentations/medicaid_disease_management_house_012605.pdf>.

168) Catalla, Robert, F.S. Goldstein, and C. Farthing. “The Disease Management Initiative – A Novel Approach in the Care of Patient with HIV/AIDS.” A poster presentation at the United States conference on AIDS, September 2001.

169) Liu, G. G., D. Ying, R. Lyu. “Economic Costs of HIV Infection: An Employer’s Perspective.” *The European Journal of Health Economics*, 3 (2002): 226-34.

170) Burton, Wayne and Catherine M. Connerty. “Worksite-based Diabetes Disease Management Program.” *Disease Management*, 5 (2002): 1-8.

171) Penttila, Chris. “An Ounce of Prevention…” *Entrepreneur Magazine*. January 2003. <http://www.findarticles.com/p/articles/mi_m0DTI/is_1_31/ai_n13470627>.

172) AcademyHealth. "Ensuring Quality Health Plans: A Purchaser's Toolkit for Using Incentives." 14, <http://www.academyhealth.org/nhcpi/healthplanstoolkit.pdf>.

173) HospitalConnect. "The Impact of the Proposed HIPAA Privacy Rule on the Hospital Industry." December 2000. <http://www.hospitalconnect.com/aha/key_issues/hipaa/content/FCGDecember2000.pdf>.

174) HospitalConnect. "Report on the Impacts of the HIPAA Final Privacy Rule on Hospitals." March 2001. <http://www.hospitalconnect.com/aha/key_issues/hipaa/content/FCGMarch2001.doc>.

175) Center to Advance Palliative Care (CAPC). "Utilization Review," *CPAC Manual*, 20 February 2002, <http://64.85.16.230/educate/content/development/utilizationreview.html>.

176) Case Management Society of America (CMSA). "CMSA Definition and Philosophy." <http://www.cmsa.org/AboutUs/CMDefinition.aspx>.

177) Case Management Society of America. "Strategic Vision." <http://www.cmsa.org/PDF/StrategicVision.pdf>.

178) Wang, Samuel, Blackford Middleton, Lisa A. Prosser, Christiana G. Bardon, Cynthia D. Spurr, Patricia J. Carchidi, Anne F. Kittlera, Robert C. Goldszer, David G. Fairchild, Andrew J. Sussman, Gilad J. Kuperman, and David W. Bates. "A Cost-benefit Analysis of Electronic Medical Records in Primary Care." *The American Journal of Medicine*, 114 (2003): 397-403.

179) Franzini, Luisa, Elena Marks, Polly F. Cromwell, Jan Risser, Laurie McGill, Christine Markham, Beatrice Selwyn, and Carrie Shapiro. "Projected Economic Costs Due to Health Consequences of Teenagers' Loss of Confidentiality in Obtaining Reproductive Health Care Services in Texas." *Archives of Pediatrics & Adolescent Medicine,* 158 (2004): 1140-46.

180) Competitive Enterprise Institute. "With a Grain of Salt: What Consumer Privacy Surveys Don't Tell Us." June 2001. 1-18, <http://www.cei.org/PDFs/with_a_grain_of_salt.pdf>.

181) Ecommerce Times. "Top Dog Oracle Losing Database Market Share." 11 March 2003. <http://www.ecommercetimes.com/story/20968.html>.

182) Database Journal. "SQL Server 2005 Security - Part 3 Encryption." 22 February 2005. <http://www.databasejournal.com/features/mssql/article.php/3483931>.

183) Still, Jared. "Data Obfuscation and Encryption." <http://www.cybcon.com/~jkstill/util/encryption/data_obfuscation_and_encryption.html>.

184) IBM. "Cost of Encryption for DB2." *IBM DB2 Tools.* <http://publib.boulder.ibm.com/infocenter/dzichelp/index.jsp?topic=/com.ibm.imstools.deu.doc.ug/cost.htm>.

185) University of Pittsburgh, Health Sciences Library System. "De-Identification Tool for Patient Records Used in Clinical Research." <http://www.hsls.pitt.edu/about/news/hslsupdate/2004/june/iim_de_id/>.

186) University of Pittsburgh. "Clinical Research Informatics Service (CRIS), Policies and Procedures." 22 July 2005. <http://www.clinicalresearch.pitt.edu/irs/services/CRIS/Policy.cfm>.

187) Goodwin, Linda, Jonathan C. Prather. "Protecting Patient Privacy in Clinical Data Mining." *Journal of Healthcare Information Management,* 16 (2002): 62-67.

188) Bureau of Labor Statistics. "Occupational Employment and Wages." 19 April 2005. <http://www.bls.gov/oes/2003/november/oes151031.htm>.

189) Bureau of Labor Statistics. "Employment Cost Trends." <http://data.bls.gov/cgi-bin/surveymost?cc> (selecting civilian, all workers, total benefits).

190) Bureau of Labor Statistics. "Research and Development in the Physical, Engineering, and Life Sciences." 24 November 2004. <http://www.bls.gov/oes/2003/november/naics5_541710.htm#b13-0000>.

191) Centers for Medicare and Medicaid services. "Billing Procedures." *Hospital Manual,* 16 September 2004. <http://www.cms.hhs.gov/manuals/10_hospital/ho460.asp>.

192) State of California, Medi-cal. "UB-92 Completion: Inpatient Services." September 2003. <http://files.medi-cal.ca.gov/pubsdoco/publications/masters-MTP/Part2/ubcompip_i00.doc>.

193) Wisconsin Department of Health and Family Services. "UB-92 (CMS 1450) Claim Form Instructions for Personal Care Services." <http://dhfs.wisconsin.gov/medicaid3/updates/2003/2003pdfs/2003-69att4.pdf>.

194) US Department of Health and Human Services, Office for Civil Rights. "Standards for Privacy of Individually Identifiable Health Information." August 2003. <http://www.hhs.gov/ocr/combinedregtext.pdf>.

195) Sweeney, Latanya. *Computational Disclosure Control: A Primer on Data Privacy Protection.* Cambridge, MA: Massachusetts Institute of Technology, 2001.

196) University of Miami. "De-identified Health Information (HIPAA)," *Privacy/data Protection Project.* <http://privacy.med.miami.edu/glossary/xd_deidentified_health_info.htm>.

197) DxCG. "Technical Requirements." <http://www.dxcg.com/uses/index.html>.

198) Quintiles.com. "Understanding the Impact of HIPAA on Clinical Research." June 26-27. 2003. <http://www.quintiles.com/NR/rdonlyres/e7n4reuv4qqzqtjpdpzirnsyzbhmygzs4uijtkagmy5gie5dvu5gvyhkmme5dwrbbcmoffkiiteejc/JBeachBarnett03.pdf>.

199) Winkler, William. "The State of Record Linkage and Current Research Problems." <http://www.census.gov/srd/papers/pdf/rr99-04.pdf>.

200) Scheuren, Fritz and William Winkler. "Regression Analysis of Data Files That Are Computer Matched – Part I." *National Research Council. Record Linkage Techniques – 1997: Proceedings of an International Workshop and Exposition,* 106-125. Washington, DC: National Academy Press, 1999.

201) Grannis, Shaun, J. Marc Overhage, and Clement McDonald. "Real World Performance of Approximate String Comparators for Use in Patient Matching." 2004, 43-7. <http://www.cs.mun.ca/~harold/Courses/Old/CS6772.F04/Diary/5604Grannis.pdf>.

202) Roos, L. L. and A. Wajda. "Record Linkage Strategies: Part I: Estimating Information and Evaluating Approaches." *Methods of Information in Medicine,* 30 (1991): 117-123.

203) Centers for Disease Control and Prevention. "Summary Health Statistics for US Adults: National Health Interview Survey, 2003." July 2005. <http://www.cdc.gov/nchs/data/series/sr_10/sr10_225.pdf>.

204) Medical Expenditures Panel Survey. "2002 Compendium of Tables – Household Medical Expenditures." 23 December 2004. <http://www.meps.ahrq.gov/mepsnet/tc/TC15.asp?_SERVICE=MEPSSocket1&_PROGRAM=MEPSPGM.TC.SAS&File=HCFY2002&Table=HCFY2002%5FPLEXP%5F%40&VAR1=AGE&VAR2=SEX&VAR3=RACETHNX&VAR4=INSURCOV&VAR5=POVCAT02&VAR6=MSA&VAR7=REGION&VAR8=HEALTH&VARO1=5+17+44+64&VARO2=1&VARO3=1&VARO4=1&VARO5=1&VARO6=1&VARO7=1&VARO8=1&_Debug>.

205) America's Health Insurance Plans. <http://www.ahip.org/>.

206) DxCG, *DxCG RiskSmart Stand Alone User Guide Version 2.0.1*, January 2005.

207) Russell, Deborah and G.T. Gangemi, Sr. "Chapter 3: Computer System Security and Access Controls." *Computer Security Basics*, <http://www.oreilly.com/catalog/csb/chapter/ch03.html>.

208) Menezes, Alfred, Paul C. van Oorschot and Scott A. Vanstone. "Chapter 1." *Handbook of Applied Cryptography*. 2001. 1-48, <http://www.cacr.math.uwaterloo.ca/hac/about/chap1.pdf>.

209) RSA Security. "What Is a Hash Function?" <http://www.rsasecurity.com/rsalabs/node.asp?id=2176>.

210) Fair, Martha. "Recent Developments at Statistics Canada in the Linking of Complex Health Files." <http://www.fcsm.gov/99papers/fair.pdf>.

211) Meyer, Scott. "Using Microsoft Access to Perform Exact Record Linkages." 1997. 280-286, <http://www.fcsm.gov/working-papers/smeyer.pdf>.

212) Liu, Shiliang and Shi Wu Wen. "Development of Record Linkage of Hospital Discharge Data for the Study of Neonatal Readmission." 2000. <http://www.phac-aspc.gc.ca/publicat/cdic-mcc/20-2/c_e.html>.

213) Winkler, William. "Matching and Record Linkage." 1997. 374-403, <http://www.fcsm.gov/working-papers/wwinkler.pdf>.

214) Fellegi, Ivan and A.B. Sunter. "A Theory for Record Linkage." *Journal of the American Statistical Association*, 64 (1969): 1183-1210.

215) Blakely, Tony and Clare Salmond. "Probabilistic Record Linkage and a Method to Calculate the Positive Predictive Value." *International Journal of Epidemiology*, 31 (2002): 1246-52.

216) Fair, Martha and Patricia Whitridge. "Tutorial on Record Linkage Slides Presentation." 1997. 457-479, <http://www.fcsm.gov/working-papers/mfair-tutorial.pdf>.

217) Grannis, Shaun, J.M. Overhage, S. Hui, and C.J. McDonald. "Analysis of a Probabilistic Record Linkage Technique Without Human Review" (American Medical Informatics Association 2003 Symposium Proceedings).

218) Quantin, C., C. Binquet, F.A. Allaert, B. Cornet, R. Pattisina, G. Leteuff, C. Ferdynus, and J.B. Gouyon. "Decision Analysis for the Assessment of a Record Linkage Procedure." *Methods of Information in Medicine*, 44 (2005): 72-79.

219) White, David. "A Review of the Statistics of Record Linkage for Genealogical Research." 1997. 362-73, <http://www.fcsm.gov/working-papers/dwhite.pdf>.

220) Damerau, Fred. "A Technique for Computer Detection and Correction of Spelling Errors." *Communications of the ACM*, 7 (1964): 171-6.

221) Quantin, Catherine, C. Binquet, K. Bourquard, R. Pattisina, B. Gouyon-Cornet, C. Ferdynus, J.B. Gouyon, and F.A. Allaert. "A Peculiar Aspect of Patients' Safety: The Discriminating Power of Identifiers for Record Linkage." *Studies in Health Technology and Informatics*, 103 (2004): 400-406.

222) Grannis, Shaun, J.M. Overhage, and C.J. McDonald. "Analysis of Identifier Performance Using a Deterministic Linkage Algorithm" (Proceedings of the AMIA 202 Annual Symposium).

223) Cook, L. J., L.M. Olson, and J.M. Dean. "Probabilistic Records Linkage: Relationships Between File Sizes, Identifiers, and Match Weights." *Methods of Information in Medicine*, 40 (2001): 196-203.

224) Yancey, William. "An Adaptive String Comparator for Record Linkage." 19 February 2004. 1-24, <http://www.census.gov/srd/papers/pdf/rrs2004-02.pdf>.

225) Hwang, Min-Shiang and Wei-Pang Yang. "A Two-phase Encryption Scheme for Enhancing Database Security." *Journal of Systems Software*, 31 (1995): 257-265.

226) Davida, George I. and David Wells. "A Database Encryption System with Subkeys." *ACM Transactions on Database Systems*, 6 (1981): 312-328.

227) Schneier, Bruce. "Security Pitfalls in Cryptographic Design." *Information Management & Computer Security*, 6 (1998): 133-137.

228) RSA Security. "What is Cryptanalysis?" *RSA Laboratories' Frequently Asked Questions about Today's Cryptography*. <http://www.rsasecurity.com/rsalabs/node.asp?id=2200>.

229) Menezes, Alfred, Paul C. van Oorschot, and Scott A. Vanstone. "Chapter 9." *Handbook of Applied Cryptography*. 2001. 321-83, <http://www.cacr.math.uwaterloo.ca/hac/about/chap9.pdf>.

230) Freesoft.org. "Connected: An Internet Encyclopedia: Key Management." <http://www.freesoft.org/CIE/Topics/138.htm>.

231) Goldwasser, Shafi and Mihir Bellare. *Lecture Notes on Cryptography*. Cambridge, MA: Massachusetts Institute of Technology, August 2001.

232) Social Security Administration. "Is There Any Significance to the Numbers Assigned in the Social Security Number?" 2003. <http://ssa-custhelp.ssa.gov/cgi-bin/ssa.cfg/php/enduser/std_adp.php?p_sid=mtTFC74h&p_lva=&p_faqid=87&p_created=955483216&p_sp=cF9zcmNoPSZwX2dyaWRzb3J0PSZwX3Jvd19jbnQ9NjImcF9jYXRfbHZsMT0xNiZwX3BhZ2U9MQ**&p_li>.

233) Social Security Administration. "Social Security Number Allocations." <http://www.socialsecurity.gov/foia/stateweb.html>.

234) Ritter, Terry. "Brute Force Attack." *Ritter's Crypto Glossary and Dictionary of Technical Cryptography*, 12 March 2004. <http://www.ciphersbyritter.com/GLOSSARY.HTM#BruteForceAttack>.

235) Lindell, Yehuda. "Pseudorandomness and Private-key Encryption Schemes." *Introduction to Cryptography*, <www.cs.biu.ac.il/~lindell/89-656/lecture03-89-656.ps>.

236) Wikipedia. "Key (Cryptography)." 14 July 2005. <http://en.wikipedia.org/wiki/Key_(cryptography)>.

237) Menezes, Alfred, Paul C. van Oorschot, and Scott A. Vanstone. "Chapter 7." *Handbook of Applied Cryptography*. 2001. 223-282, <http://www.cacr.math.uwaterloo.ca/hac/about/chap7.pdf>.

238) Menezes, Alfred, Paul C. van Oorschot, and Scott A. Vanstone. "Chapter 8." *Handbook of Applied Cryptography*. 2001. 283-319, <http://www.cacr.math.uwaterloo.ca/hac/about/chap8.pdf>.

239) Touloumtzis, Mike. "Re: (AES) Loopback Crypto Questions." 11 July 2001. <http://mail.nl.linux.org/linux-crypto/2001-07/msg00150.html>.

240) Reavis, Jim. "Feature: Goodbye DES, Hello AES." *Networkworld*, July 30, 2001, <http://www.networkworld.com/research/2001/0730feat2.html>.

241) Candan, K.S., Sushil Jajodia, and V.S. Subrahmanian. "Secure Mediated Databases" (Proceedings—International Conference on Data Engineering, 1996).

242) Agrawal, Rakesh, Alexandre Evfimievski, and Ramakrishnan Srikant. "Information Sharing Across Private Databases" (Association for Computing Machinery, Special Interest Group on Management of Data, June 9-12, 2003).

243) Benaloh, Josh Cohen. "Cryptographic Capsules: A Disjunctive Primitive for Interactive Protocols" (Proceedings on Advances in Cryptology – CRYPTO '86, Santa Barbara, California).

244) Fagin, Ronald, Moni Naor, and Peter Winkler. "Comparing Information without Leaking It." *Communications of the ACM*, 39 (1996): 77-85.

245) Rivest, Ronald, L. Adleman, and M. L. Dertouzos. "On Data Banks and Privacy Homomorphisms." In *Foundations of Secure Computation*, edited by R.A. DeMillo, 169-177. New York: Academic Press, 1978.

246) Blakley, G.R., and Catherine Meadows. "A Database Encryption Scheme which Allows the Computation of Statistics using Encrypted Data" (IEEE Symposium on Security and Privacy, April 22-24, 1985).

247) Quantin, Catherine, François-André Allaert, and Liliane Dusserre. "Anonymous Statistical Methods versus Cryptographic Methods in Epidemiology." *International Journal of Medical Informatics*, 60 (2000): 177-83.

248) Domingo-Ferrer, Josep. "A Provably Secure Additive and Multiplicative Privacy Homomorphism." In *Lecture Notes in Computer Science 2433*, edited by AH Chan, 471-483. London: Springer-Verlag, 2002.

249) Naor, Moni and Kobbi Nissim. "Communication Preserving Protocols for Secure Function Evaluation." <http://www.wisdom.weizmann.ac.il/~kobbi/papers/sfe_proc.ps>.

250) Sandhu, Ravi and Pierangela Samarati. "Access Control: Principles and Practice." *IEEE Communications Magazine*, 32 (1994): 40-8.

251) Date, C.J. *Introduction to Database Systems, Sixth Edition*. Reading, MA: Addison-Wesley Publishing Company, 1995.

252) Goldwasser, Shafi. "Lecture 7: Zero Knowledge" (handout given at lecture at MIT, August 2001).

253) Chaum, David, Claude Crepeau, and Ivan Damgard. "Multiparty Unconditionally Secure Protocols" (Proceedings of the 20th Symposium on the Theory of Computing, 1988).

254) Chaum, David, Ivan Damgard, and Jeroen van de Graaf. "Multiparty Computations Ensuring Privacy of Each Party's Input and Correctness of the Result." *Lecture Notes in Computer Science*, 293 (1988): 87-119.

255) Zero Knowledge Systems. "Private Credentials." November 2000. 1-25, <http://osiris.978.org/~brianr/crypto-research/anon/www.freedom.net/products/whitepapers/credsnew.pdf>.

256) Domingo-Ferrer, Josep. "Advances in Inference Control in Statistical Databases: An Overview." <http://neon.vb.cbs.nl/casc/overview.pdf>.

257) Song, Dawn Xiaodong, David Wagner, and Adrian Perrig. "Practical Techniques for Searches on Encrypted Data" (IEEE Symposium on Security and Privacy, 2000).

258) Wikipedia. "Deterministic Encryption." 20 July 2005. <http://en.wikipedia.org/wiki/Deterministic_encryption>.

259) Du, Wenliang and Mikhail Atallah. "Protocols for Secure Remote Database Access with Approximate Matching" (7th ACM Conference on Computer and Communications Security (ACMCCS 2000), The First Workshop on Security and Privacy in E-Commerce).

260) Wikipedia. "Probabilistic Encryption." 3 September 2005. <http://en.wikipedia.org/wiki/Probabilistic_encryption>.

261) Churches, Tim and Peter Christen. "Some Methods for Blindfolded Record Linkage." *BMC Medical Informatics and Decision Making*, 4 (2004): 1-17, <http://www.biomedcentral.com/content/pdf/1472-6947-4-9.pdf>.

262) Sam's String Metrics. "Dice's Coefficient." <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#dice>.

263) Feige, Uri, Joe Kilian, and Moni Naor. "A Minimal Model for Secure Computation." 1-15, <http://www.wisdom.weizmann.ac.il/~naor/PAPERS/fkn.pdf>.

264) DxCG. <http://www.dxcg.com/>.

265) DxCG. *About the Company*. <http://www.dxcg.com/about/index.html>.

266) MySQL. <http://www.mysql.com/>.

267) DxCG, *RiskSmart Models and Methodologies* (2002).

268) Ingenix Corporation. "Identification and Management of High Risk Patients Using a Claims-based Predictive Model." <http://www.ingenix.com/esg/resourceCenter/10/~dt_pyr_wp_1-03.pdf>.

269) Massachusetts Institute of Technology. "De-identified Data," *Committee on the Use of Humans as Experimental Subjects*. 30 November 2004. <http://web.mit.edu/committees/couhes/definitions.shtml#De-identifiedData>.

270) Symmetry. "A Comparative Analysis of Claims-based Methods of Health Risk Assessment for Commercial Populations." 24 May 2002. <http://www.symmetry-health.com/SOAStudy.pdf>.

271) Ash, Arlene, Yang Zhao, Randall Ellis, and Marilyn Schlein Kramer. "Finding Future High-cost Cases: Comparing Prior Cost versus Diagnosis-based Methods." *Health Services Research*, 36 (2001): 194-206.

272) Tatlow, James, John D. Clapp, and Melinda M. Hohman. "The Relationship between the Geographic Density of Alcohol Outlets and Alcohol-related Hospital

Admissions in San Diego County." *Journal of Community Health*, 25 (2000): 79-88.

273) Lander University. Class notes for NURS 416. <http://www.lander.edu/bfreese/416%20Notes%20Ch%2021.doc>.

274) Nanchahal, Kiran, Punam Mangtani, Mark Alston, and Isabel dos Santos Silva. "Development and Validation of a Computerized South Asian Names and Group Recognition Algorithm (SANGRA) for Use in British Health-related Studies." *Journal of Public Health Medicine*, 23 (2001): 278-85.

275) Selby, Joe V., Andrew J. Karter, Lynn M. Ackerson, Assiamira Ferrara, and Jennifer Liu. "Developing a Prediction Rule from Automated Clinical Databases to Identify High-risk Patients in a Large Population with Diabetes." *Diabetes Care*, 24 (2001): 1547-55.

276) Forman, Samuel, Matthew Kelliher, and Gary Wood. "Clinical Improvement with Bottom-line Impact: Custom Care Planning for Patients with Acute and Chronic Illnesses in a Managed Care Setting." *The American Journal of Managed Care*, 3 (1997): 1039-48.

277) Lynch, John, Samuel A. Forman, Sandy Graff, and Mark C. Gunby. "High-risk Population Health Management – Achieving Improved Patient Outcomes and Near-term Financial Results." *The American Journal of Managed Care*, 6 (2000): 781-91.

278) Dove, Henry, Ian Duncan, and Arthur Robb. "A Prediction Model for Targeting Low-cost, High-risk Members of Managed Care Organizations." *The American Journal of Managed Care*, 9 (2003): 381-9.

279) Perls, Thomas and Elizabeth R. Wood. "Acute Care Costs of the Oldest Old: They Cost Less, Their Case Intensity is Less, and They Go to Nonteaching Hospitals." *Archives of Internal Medicine*, 156 (1996): 754-60.

280) Le, Chap T., Ping Liu, Bruce R. Lindgren, Kathleen A. Daly, and G. Scott Giebink. "Some Statistical Methods for Investigating the Date of Birth as a Disease Indicator." *Statistics in Medicine*, 22 (2003): 2127-35.

281) Liu, Xiaofeng, Roland Sturm, and Brian J. Cuffel. "The Impact of Prior Authorization on Outpatient Utilization in Managed Behavioral Health Plans." *Medical Care Research and Review*, 57 (2000): 182-95.

282) Palmer, K. T., M. J. Griffin, H. E. Syddall, A. Davis, B. Pannett, and D. Coggon. "Occupational Exposure to Noise and the Attributable Burden of Hearing Difficulties in Great Britain." *Occupational and Environmental Medicine*, 59 (2002): 634-9.

283) Heinze, Daniel, Mark L. Morsch, and John Holbrook. "Mining Free-text Medical Records" (Proceedings of the American Medical Informatics Association, 2001).

284) Petrou, Stavros, Ziyah Mehta, Christine Hockley, Paula Cook-Mozaffari, Jane Henderson, and Michael Goldacre. "The Impact of Preterm Birth on Hospital Inpatient Admissions and Costs during the First 5 Years of Life." *Pediatrics*, 112 (2003): 1290-7.

285) Little One Productions. "Neonatal Intensive Care." <http://www.littleoneprods.com/neonatalintensiv.html>.

286) Lee, Shoo, Douglas D. McMillan, Arne Ohlsson, Margaret Pendray, Anne Synnes, Robin Whyte, Li-Yin Chien, and Joanna Sale. "Variations in Practice and

Outcomes in the Canadian NICU Network: 1996-1997." *Pediatrics*, 106 (2000), 1070-1079, <http://pediatrics.aappublications.org/cgi/content/full/106/5/1070>.

287) Babycenter. "Preterm Labor and Birth." <http://www.babycenter.com/refcap/pregnancy/pregcomplications/1055.html>.

288) Krentz, H.B. "The High Cost of Medical Care for Patients Who Present Late (CD4 < 200 cells/uL) with HIV Infection." *HIV Medicine*, 5 (2004): 93-98.

289) "Jury Orders Hospital to Pay $250,000 for Invasion of Privacy." *Aids Policy & Law*, 15 (2000): 10.

290) Tricare. "Tricare Benefits for College Students." 26 February 2004. <http://www.tricare.osd.mil/collegestudents/TRICAREClaims.cfm>.

291) TMG Health. "Services." <http://www.tmghealth.com/services/claims.html>.

292) Interim Healthcare. "Retrospective Claims Analysis." <http://www.interimhealthcare.com/biz/CorpRetailPharma/EmpWellnessProgram/RetrospectiveClaimsAnalysis.asp>.

293) Tufts Health Plan. "Claims Procedures," *Billing Guidelines*. <http://www.tuftshealthplan.com/providers/provider.php?sec=administrative_resources&content=b_COB&rightnav=billing>.

294) Health Data Management. "Easy Access to Data Helps Insurers Make Timely Decisions." <http://www.healthdatamanagement.com/html/guide/toptech03alisting.cfm?AdvertiserID=753>.

295) Poria, Yaniv and Harmen Oppewal. "Student Preferences for Room Attributes at University Halls of Residence: An Application of the Willingness to Pay Technique." *Tourism and Hospitality Research*, 4 (2002): 116-30.

296) Tutor2u. "Sales Forecasting." <http://www.tutor2u.net/business/marketing/sales_forecasting.asp>.

297) Ford, Carol, Abigail English, and Garry Sigman. "Confidential Health Care for Adolescents: Position Paper of the Society of Adolescent Medicine." *Journal of Adolescent Health*, 35 (2004): 160-7.

298) Federal Trade Commission. "Online Profiling: A Report to Congress, Part 2 Recommendations." July 2000. <http://www.ftc.gov/os/2000/07/onlineprofiling.htm>.

299) Government of British Columbia. "Chronic Disease Management," *Ministry of Health Services*. 21 September 2005. <http://www.healthservices.gov.bc.ca/cdm/>.

300) National Institute of Standards and Technology. "Announcing the Advanced Encryption Standard." 26 November 2001. 1-47, <http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf>.

301) Feigenbaum, Joan, Michael J. Freedman, Tomas Sander, and Adam Shostack. "Economic Barriers to the Deployment of Existing Privacy Technologies." 2002. 1-3, <http://citeseer.ist.psu.edu/cache/papers/cs/26515/http:zSzzSzcs-www.cs.yale.eduzSzhomeszSzjfzSzPrivacy-Barriers-WP.pdf/feigenbaum02economic.pdf>.

302) Shaw, Thomas R. "The Moral Intensity of Privacy: An Empirical Study of Webmasters' Attitudes." *Journal of Business Ethics*, 46 (2003): 301-318.

303) Office of Technology Assessment. *Protecting Privacy in Computerized Medical Information*. Washington, DC: US Government Printing Office, 1993.

304) Robinson, James C. "The Politics of Managed Competition: Public Abuse of the Privacy Interest." *Journal of Health Politics, Policy, and Law*, 28 (2003): 341-353.

305) Christakis, Dimitri, Anne Kazak, Jeffrey Wright, Frederick Zimmerman, Alta Bassett, and Frederick Connell. "What factors are associated with achieving high continuity of care?" *Family Medicine*, 36 (2004): 55-60.

306) Weiner, Jonathan P., Stephen T. Parente, Deborah W. Garnick, Jinnet Fowles, Ann G. Lawthers, and R. Heather Palmer. "Variation in Office-Based Quality: A Claims-Based Profile of Care Provided To Medicare Patients With Diabetes." *JAMA*, 273 (1995): 1503-1508.

307) CorSolutrions. "Cancer Solutions." <http://www.corsolutions.com/programs/2.3.7_cancer.html>.

308) Health Management Corporation. "Healthy Returns Program for Low Back Pain." <http://www.choosehmc.com/LowBackPain.html>.

309) LifeMasters. "Products & Services." <http://www.lifemasters.com/corporate/prod/index.asp>.

310) Accordant. "Rheumatoid Arthritis." <http://www.accordant.net/ra.html>.

311) University of Texas Health Science Center at San Antonio. "Evaluation for Business Associates," *HIPAA Compliance Program*. 10 October 2005. <http://www.uthscsa.edu/hipaa/assoc-who.html>.

312) US Department of Health and Human Services, Office for Civil Rights. "Business Associates." 3 April 2003. <http://www.hhs.gov/ocr/hipaa/guidelines/businessassociates.rtf>.

313) The Library of Congress. "S. 768," *Thomas*. <http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:s768is.txt.pdf>.

314) The Library of Congress. "S. 1408," *Thomas*. <http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:s1408is.txt.pdf>.

315) The Library of Congress. "S. 1332," *Thomas*. <http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_cong_bills&docid=f:s1332pcs.txt.pdf>.

316) Federal Trade Commission. "Nation's Big Three Consumer Reporting Agencies Agree To Pay $2.5 Million To Settle FTC Charges of Violating Fair Credit Reporting Act." <http://www.ftc.gov/opa/2000/01/busysignal.htm>.

317) Wrongdiagnosis.com. "Prevalence and Incidence of HIV/AIDS." 2 April 2003. <http://www.wrongdiagnosis.com/h/hiv_aids/prevalence.htm>.

318) Wrongdiagnosis.com. "Prevalence and Incidence of Diabetes." 10 April 2003. <http://www.wrongdiagnosis.com/d/diabetes/prevalence.htm>.

319) Wrongdiagnosis.com. "Prevalence and Incidence of Asthma." 9 April 2003. <http://www.wrongdiagnosis.com/a/asthma/prevalence.htm>.

320) Wrongdiagnosis.com. "Prevalence and Incidence of Coronary Heart Disease." <http://www.wrongdiagnosis.com/c/coronary_heart_disease/prevalence.htm>.

321) Wrongdiagnosis.com. "Prevalence and Incidence of Congestive Heart Failure." 23 October 2003. <http://www.wrongdiagnosis.com/c/congestive_heart_failure/prevalence.htm>.

322) Joint Rehabilitation & Sports Medical Center. "Lower Back Pain." <http://www.jointrehab.com/lower_back_pain.htm>.

323) US Census Bureau. "National and State Population Estimates." <http://www.census.gov/popest/states/tables/NST-EST2005-01.xls> (Microsoft Excel file).

324) Robinson, James. "The Future of Managed Care Organization." *Health Affairs*, 18 (1999): 7-24.

325) Centers for Disease Control and Prevention. "Basic Statistics." 26 January 2006. <http://www.cdc.gov/hiv/topics/surveillance/basic.htm>.

326) Centers for Disease Control and Prevention. "Current Asthma Prevalence Percents by Age, United States: National Health Interview Survey, 2003." 9 March 2005. <http://www.cdc.gov/asthma/NHIS/2003_Table_4-1.pdf>.

327) Centers for Disease Control and Prevention. "National Diabetes Fact Sheet," *Publications and Products*. 16 November 2005. <http://www.cdc.gov/diabetes/pubs/estimates05.htm#prev>.

328) Centers for Disease Control and Prevention. "Congestive Heart Failure and Adrenergic Receptor Polymorphisms," *Genomics and Disease Prevention*. 27 November 2002. <http://www.cdc.gov/genomics/hugenet/ejournal/heartfailure.htm#2>.

329) Agency for Healthcare Research and Quality. "Chronic Stable Coronary Artery Disease (CAD): Percentage of Patients Who Were Screened for Diabetes," *National Quality Measures Clearinghouse*. 6 March 2006. <http://www.qualitymeasures.ahrq.gov/summary/summary.aspx?ss=1&doc_id=7830>.

330) Carey TS, Evans A, Hadler N, Kalsbeek W, McLaughlin C, Fryer J. "Care-seeking among individuals with chronic low back pain." *Spine*, 20 (1995): 312-7.

331) Robert Wood Johnson Foundation. "Chronic Care in America: A 21st Century Challenge." November 1996. <http://www.rwjf.org/files/publications/other/ChronicCareinAmerica.pdf>.

332) "Diagnoses of HIV/AIDS—32 States, 2000-2003," *Morbidity and Mortality Weekly Report*, 53 (2004): 1106-1110.

333) Thom, Thomas, Nancy Haase, Wayne Rosamond, Virginia Howard. "Heart Disease and Stroke Statistics – 2006 Update. *Circulation*, 113 (2006): e85-e151.