

# **Linking Data for Health Services Research**

A Framework and Guidance for Researchers

DRAFT REPORT

Contract: HHSA2902010000141

Date: 04-November-2013

# Table of Contents

<b>Chapter 1. Background and purpose</b> .....	<b>1</b>
Overview .....	1
Background.....	1
Purpose .....	4
References .....	6
<b>Chapter 2. Research environment</b> .....	<b>9</b>
Overview.....	9
<b>Computing systems and the balance between security and usability</b> .....	9
Desktop computer: high security, low usability, low cost.....	10
Central server: medium security, medium usability, moderate cost.....	10
Virtual remote desktops: high security, high usability, high cost .....	10
<b>Building the technical platform</b> .....	11
Securing the platform.....	11
Users accessing platform .....	11
Gaining access to platform .....	12
Processing power of platform .....	12
Data storage platform .....	12
<b>Securing the research environment</b> .....	13
Overview .....	13
Regulatory requirements .....	14
Identifying sensitive data .....	15
Summary of protected information.....	17
Building and implementing a security plan .....	18
Workforce training .....	19
<b>Managing risks</b> .....	20
Identifying and assessing risk .....	21
Tracking risks .....	21
Planning risk responses .....	23
Implementing risk responses .....	23
Monitoring risks .....	23
<b>Conclusions</b> .....	24
<b>Appendix 1. Procedures and processes to enhance data security</b> .....	25
<b>Chapter 3. Linkage feasibility--to link or not to link</b> .....	<b>30</b>
Overview .....	30
Evaluating linkage feasibility .....	30
Purpose and conditions for original data collection .....	34
Ownership of data .....	34

Data sharing and security concerns .....	36
Building the team .....	37
References .....	38
<b>Chapter 4. An overview of record linkage methods .....</b>	<b>39</b>
Overview .....	39
Data cleaning and standardization .....	40
Linkage methods .....	42
Deterministic linkage methods .....	43
Probabilistic linkage methods .....	44
Alternative linkage methods .....	48
Selecting a linkage method .....	48
Evaluating linkage algorithms .....	49
Validating linkage results .....	50
Final remarks .....	51
References .....	53
Appendix 2. Useful SAS functions and procedures .....	58
Appendix 3. Data linkage software packages .....	66
<b>Chapter 5. An evaluation of methods linking health registry data to insurance claims in scenarios of varying available information .....</b>	<b>68</b>
Objective .....	68
Methods .....	68
Approach overview .....	68
Data sources and patient populations .....	69
Data cleaning and standardization .....	70
Data linkage .....	70
Results .....	74
Discussion .....	80
References .....	81
Appendix 4. SEER-Medicare algorithm with partial identifiers .....	82
<b>Chapter 6. Project summary and recommendations for researchers .....</b>	<b>85</b>
Overview .....	85
Considerations for project planning .....	85
Appropriateness and feasibility of the project .....	85
Data ownership and governance .....	86
Technical environment and security .....	86
Team, skills, and expertise .....	86
Cost .....	87
Identify and evaluate available linkage keys .....	87

<b>Variable cleaning, standardization, and a common data model (Normalization)</b> .....	87
<b>Linkage approach</b> .....	88
<b>Evaluation and validation of record linkage</b> .....	89
<b>Recommendations for reporting results</b> .....	90
<b>Framework for registry-to-claims linkage</b> .....	90
<b>Project planning checklist</b> .....	91
<b>Project execution checklist</b> .....	91

# Chapter 1. Background and purpose

## Overview

Given the recent pressure to reduce the release of sensitive identifying information, the development of effective record linkage approaches for varying scenarios of data availability is critical. The objective of this report is to present a conceptual framework and instructional information that scientifically describe the strengths and limitations of different approaches to record linkage of registries to other data sources. This chapter presents the context and motivation for the work detailed in subsequent chapters of this report. Specifically, it describes the need for data linkages in general, and registry-to-health insurance claims linkages specifically in the context of comparative effectiveness research (CER) and the environments in which CER is conducted.

## Background

Randomized controlled trials (RCTs) remain the gold standard for assessing intervention efficacy; however, RCTs are not always feasible or sufficiently timely. Perhaps more importantly, RCT results often cannot be generalized due to a lack of inclusion of “real world” combinations of interventions and heterogeneous patients.(Clancy and Slutsky 2007; Office 2007; Smith 2007; 2009) With recent advances in information technology, data, and statistical methods, there is tremendous promise in leveraging ever-growing repositories of secondary data to support comparative effectiveness and public health research.(Institute of Medicine 2009; Sox and Greenfield 2009; VanLare, Conway et al. 2010) Such research can help fill the knowledge gaps unaddressed by RCTs and extend the utility of current data investments to examine important research questions and health care programs. Because of this, the Institute of Medicine and others have strongly advocated for comparative effectiveness research (CER) using secondary data from sources such as health registries, administrative claims, and electronic health records.(Bloomrosen M 2008; Centers for Disease Control and Prevention 2010; 2011; Sturmer, Jonsson Funk et al. 2011)

Although these secondary data sources have many strengths, they also have a number of limitations that must be acknowledged. For example, because they are often collected for non-research purposes, secondary datasets do not have the benefit of randomization to control for the range of confounders associated with treatment, outcomes, and corresponding bias that threaten study validity. (Howe 1998;

Blakely and Salmond 2002; Bohensky, Jolley et al. 2010) Moreover, individual datasets are often limited in scope, which, in turn, limits their utility in addressing important questions in a comprehensive manner. These limitations can be overcome by linking data from multiple sources such as health registries and surveillance systems, administrative data, clinical information systems, and other sources. (Lipscomb, Gotay et al. 2005; Gliklich and Dreyer 2007; Bloomrosen and Detmer 2008; Brookhart, Sturmer et al. 2010) Importantly, linking a registry with external data can facilitate case and control group identification, improve measurement of risk factors and outcomes, and allow passive follow-up of study participants. (Mortensen 1995; Howe 1998; Warren, Feuer et al. 1999; Jutte, Roos et al. 2011) Linkages can also be used to refine and validate measures created using claims data or other registries (Setoguchi, Solomon et al. 2007; Hummler and Poets 2011; Li, Glynn et al. 2011) or to adjust for unmeasured confounding by using supplemental data about a subset of observations. (Sturmer, Schneeweiss et al. 2005) However, data linkage from multiple sources in support of CER and public health research continues to face several challenges.

Data linkage is the process of pairing observations from two or more files and identifying the pairs that belong to the same entity. (Winglee, Valliant et al. 2005) A common form of linkage involves pairing the information on the same person from two datasets. However, linkage errors can arise from multiple points when data sources are inconsistent in capturing the same person, that person's records do not link due to missing or inaccurate data in one or more files, or different people are erroneously linked for the same reasons. (Bohensky, Jolley et al. 2010) The causes of these errors are many, including poor data quality, data that uses different systems for coding the linking variables in their records, dataset size and inadequate capacity of linking software, and complexity of data linking systems that have substantial learning curves, among others. These factors can be compounded by recent changes that restrict access to familiar linking variables that have been a mainstay of record linkage systems to-date.

For public health and health care delivery settings, concerns regarding privacy, confidentiality, and safety have led to increasingly stringent policies and regulations governing the collection, use, and transfer of personally identifying information. Perhaps the most well-known of these is the Health Insurance Portability and Accountability Act (HIPAA), which has been widely decried by healthcare programs and researchers as exemplifying the law of unintended consequences. Intended to protect patients from misuse of their personal information, the law has been associated with a substantially greater burden for both health care providers and researchers, and even indicted as contributory to

patient deaths and research bias that may misinform and thus diminish the quality of future health care delivery. (Kulynych and Korn 2002; Kulynych and Korn 2003; Salem and Pauker 2003; Dracup and Bryan-Brown 2004; Pentecost 2004; Beebe, Ziegenfuss et al. 2011)

Of particular importance, privacy and confidentiality concerns and increasingly restrictive policies governing protected health information (PHI) severely limit access to unique identifiers (e.g., Social Security Numbers [SSNs], names), traditional mainstays of quality data linkages involving person-level data. (Fellegi I 1969; Safran, Bloomrosen et al. 2007; Bradley, Penberthy et al. 2010) Indeed, a recent consensus report by the Institute of Medicine presents that HIPAA – passed with the intention of protecting patient privacy – does not protect privacy as well as it should, and agrees with assertions that HIPAA impedes important research;(Institute of Medicine 2009) however, national policy makers continue to rely on it. Therefore, access to the familiar variables that have been the mainstay of data linkages, such as name, SSN, and other PHI, are increasingly restricted, and many programs have even stopped collecting some PHI elements.

In addition to concerns about availability of unique identifiers across datasets, there are several notable concerns regarding the datasets themselves. Because registry studies rely on external data sources and typically are not randomized, they are subject to multiple validity threats, such as confounding by indication (treatment assignment influenced by risk for the outcome), selection bias (due to inability to define an appropriate control group), and misclassification of exposures or outcomes. Additionally, linkage errors can systematically bias effect estimates (Howe 1998; Blakely and Salmond 2002; Bohensky, Jolley et al. 2010). While researchers have continued to develop new methods for deterministic and probabilistic linkage,(Blakely and Salmond 2002; Hammill, Hernandez et al. 2009; Tromp, Ravelli et al. 2011) alternative methods have not been tested thoroughly and comprehensive guidelines are lacking. (Blakely and Salmond 2002; Hammill, Hernandez et al. 2009; Tromp, Ravelli et al. 2011)

In light of these challenges, there is an urgent need to develop greater knowledge of how to perform reliable, valid linkages to support current programs and leverage the investment already made in them to examine important comparative effectiveness research (CER) and public health questions. Without such gains in knowledge and capacity, healthcare programs will miss substantial opportunities to use vastly expanding data resources to improve the public's health, and many stakeholders will remain skeptical of

clinical or policy decisions based on incomplete data.(Keyhani, Woodward et al. 2010; Sox 2010; Sox, Helfand et al. 2010)

## **Purpose**

This report serves as a conceptual framework describing the spectrum of activity and requirements for high-quality data linkage in the context of CER, including the strengths and limitations of different approaches to record linkage of registries to other data sources. It serves as an instructional guide for researchers designing new CER studies using patient registries linked with other secondary data sources. Through this report, we provide an overview of registry-to-administrative claims linkage, including considerations for researchers, data managers, information technology managers, and other stakeholders that are likely to be involved in the process of data linkage. We also provide a framework for data linkage and results from our own application of the framework to a real-world data linkage problem.

This report is informed by practical insight developed by researchers affiliated with the AHRQ DEcIDE CER Consortium who acquire, maintain, and link sensitive secondary data for purposes of supporting comparative effectiveness and outcomes research studies. Consistent with the goals of the Agency for Healthcare Research and Quality (AHRQ) and the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Network, this report will enhance the ability of CER/Patient Centered Outcomes Research (PCOR) to inform consumers, clinicians, policymakers, and other healthcare decision-makers.

Record linkage in support of CER is more than simply joining two datasets. Rather, it is an extensive process that involves many steps and collaborations between multiple partners. Researchers hoping to link data sources should be aware of the multiple technical, legal, and data-management challenges and considerations before embarking on these efforts. We describe each chapter's primary focus below.

*Chapter 2* describes the components of a secure research environment. It is set in the context of the importance of building trust among data/research partners and includes technical and administrative guides to establish a secure computing platform enabling sophisticated CER research. Additionally it discusses approaches to secure confidentiality, integrity, and availability of data in order to maintain compliance with state and federal regulations.

*Chapter 3* describes general considerations when planning to link data. The focus spans issues pertaining to requesting, receiving and managing data from different sources, Data Use Agreement/Contracts, issues of data ownership, and deciding whether data linkage is feasible and appropriate.

*Chapter 4* guides the reader through an overview of data linkage methods in an effort to document a set of best practices for conducting linkages with optimal validity and reliability. It begins with issues pertaining to the quality of the available linkage keys, specifically completeness, overlapping information, and commonly observed idiosyncrasies. It focuses on the creation of a common data model via techniques for cleaning and standardizing the linkage keys before linkage. Next, an overview of data linkage methods is provided, including a detailed summary of deterministic and probabilistic linkage methods as well as techniques for evaluating the quality of the linkage. Finally, this chapter includes a set of appendices that contain 1) lists and characteristics of several open source and commercial software packages for data linkage, 2) sample SAS for data preparation and linkage procedures, and 3) recommended readings for those interested in learning more about alternative methods.

In *Chapter 5*, an approach to linking registry data to administrative claims is developed using the methods discussed in Chapter 4. Our approach involves four components: 1) employment of a gold standard, 2a) an evaluation of deterministic approaches, 2b) an evaluation of a deterministic approach using encryption, and (3) probabilistic approaches, each applied in varying scenarios of data availability to ascertain optimal approaches in given scenarios. In step 2b, a deterministic approach using encryption, we simulate a scenario of restrictions on identifier release to researchers. Given the exceptionally limited availability of practical empirical examples researchers can use to inform their own data linkages, this examination will articulate much-needed and specific in-depth examples of the steps researchers may take and what they may expect to find given their unique scenarios of data availability and data quality.

Finally, *Chapter 6* summarizes the report overall, and provides specific recommendations for researchers who plan to undertake a data linkage project. We also provide information, including a checklist, for researchers to use in both the project-planning phase and the project execution phase.

## References

1. Clancy, C.M. and J.R. Slutsky, *Commentary: a progress report on AHRQ's Effective Health Care Program*. Health Serv Res, 2007. 42(5): p. xi-xix.
2. *Institute of Medicine. Initial National Priorities for Comparative Effectiveness Research*. Washington DC: National Academics Press, 2009.
3. Office, C.B., *Research on the Comparative Effectiveness of Medical Treatments: Issues and Options for an Expanded Federal Role*. Pub. No. 2975 Washington DC, 2007.
4. Smith, S., *Preface*. Medical Care, 2007. 45(10 Suppl 2): p. S1-S2.
5. Institute of Medicine, *Initial National Priorities for Comparative Effectiveness Research*2009, Washington DC: National Academies Press.
6. Sox, H.C. and S. Greenfield, *Comparative effectiveness research: a report from the Institute of Medicine*. Ann Intern Med, 2009. 151(3): p. 203-5.
7. VanLare, J.M., P.H. Conway, and H.C. Sox, *Five next steps for a new national program for comparative-effectiveness research*. N Engl J Med, 2010. 362(11): p. 970-3.
8. Bloomrosen M, D.D., *Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association*. J Am Med Inform Assoc, 2008. 15(6): p. 715-722.
9. Centers for Disease Control and Prevention, *FOA: Enhancing Cancer Registry Data for Comparative Effectiveness Research*2010, Atlanta, GA: CDC.
10. Sturmer, T., et al., *Nonexperimental Comparative Effectiveness Research Using Linked Healthcare Databases*. Epidemiology, 2011. 22(3): p. 298-301.
11. *Institute of Medicine. Engineering a Learning Healthcare System: A Look at the Future: Workshop Summary*. Washington DC: National Academics Press, 2011.
12. Blakely, T. and C. Salmond, *Probabilistic record linkage and a method to calculate the positive predictive value*. Int J Epidemiol, 2002. 31(6): p. 1246-52.
13. Bohensky, M.A., et al., *Data linkage: a powerful research tool with potential problems*. BMC Health Serv Res, 2010. 10: p. 346.
14. Howe, G.R., *Use of computerized record linkage in cohort studies*. Epidemiol Rev, 1998. 20(1): p. 112-21.
15. Lipscomb, J., C. Gotay, and C. Snyder, *Outcomes Assessment in Cancer: Measures, Methods, and Applications*2005, Cambridge: Cambridge University Press.

16. Brookhart, M.A., et al., *Confounding control in healthcare database research: challenges and potential approaches*. Med Care, 2010. 48(6 Suppl): p. S114-20.
17. Gliklich, R. and N. Dreyer, eds. *Registries for Evaluating Patient Outcomes: A User's Guide (AHRQ Publication No. 07-EHC001-1)*. 2007, Agency for Healthcare Research and Quality: Rockville, MD.
18. Bloomrosen, M. and D. Detmer, *Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association*. J Am Med Inform Assoc, 2008. 15(6): p. 715-22.
19. Jutte, D.P., L.L. Roos, and M.D. Brownell, *Administrative record linkage as a tool for public health research*. Annu Rev Public Health, 2011. 32: p. 91-108.
20. Mortensen, P.B., *The untapped potential of case registers and record-linkage studies in psychiatric epidemiology*. Epidemiol Rev, 1995. 17(1): p. 205-9.
21. Warren, J.L., et al., *Use of Medicare hospital and physician data to assess breast cancer incidence*. Med Care, 1999. 37(5): p. 445-56.
22. Hummler, H.D. and C. Poets, *[Mortality of extremely low birthweight infants - large differences between quality assurance data and the national birth/death registry]*. Z Geburtshilfe Neonatol, 2011. 215(1): p. 10-7.
23. Li, Q., et al., *Validity of claims-based definitions of left ventricular systolic dysfunction in Medicare patients*. Pharmacoepidemiol Drug Saf, 2011. 20(7): p. 700-8.
24. Setoguchi, S., et al., *Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data*. Cancer Causes Control, 2007. 18(5): p. 561-9.
25. Sturmer, T., et al., *Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration*. Am J Epidemiol, 2005. 162(3): p. 279-89.
26. Winglee, M., R. Valliant, and F. Schuren, *A case study in record linkage*. Survey Methodology, 2005. 31(1): p. 3-11.
27. Pentecost, M.J., *HIPAA and the law of unintended consequences*. J Am Coll Radiol, 2004. 1(3): p. 164-5.
28. Dracup, K. and C.W. Bryan-Brown, *The law of unintended consequences*. Am J Crit Care, 2004. 13(2): p. 97-9.

29. Kulynych, J. and D. Korn, *The new HIPAA (Health Insurance Portability and Accountability Act of 1996) Medical Privacy Rule: help or hindrance for clinical research?* *Circulation*, 2003. 108(8): p. 912-4.
30. Salem, D.N. and S.G. Pauker, *The adverse effects of HIPAA on patient care.* *N Engl J Med*, 2003. 349(3): p. 309.
31. Kulynych, J. and D. Korn, *The new federal medical-privacy rule.* *N Engl J Med*, 2002. 347(15): p. 1133-4.
32. Beebe, T.J., et al., *Health Insurance Portability and Accountability Act (HIPAA) Authorization and Survey Nonresponse Bias.* *Medical care*, 2011.
33. Bradley, C.J., et al., *Health services research and data linkages: issues, methods, and directions for the future.* *Health Serv Res*, 2010. 45(5 Pt 2): p. 1468-88.
34. Fellegi I, S.A., *A Theory for Data Linkage.* *Journal of the American Statistical Association*, 1969. 64(328): p. 1183-1210.
35. Safran, C., et al., *Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper.* *J Am Med Inform Assoc*, 2007. 14(1): p. 1-9.
36. Institute of Medicine, *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research* 2009, Washington DC: National Academies Press.
37. Hammill, B.G., et al., *Linking inpatient clinical registry data to Medicare claims data using indirect identifiers.* *Am Heart J*, 2009. 157(6): p. 995-1000.
38. Tromp, M., et al., *Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage.* *J Clin Epidemiol*, 2011. 64(5): p. 565-72.
39. Keyhani, S., M. Woodward, and A.D. Federman, *Physician views on the use of comparative effectiveness research: a national survey.* *Ann Intern Med*, 2010. 153(8): p. 551-2.
40. Sox, H.C., *Comparative effectiveness research: a progress report.* *Ann Intern Med*, 2010. 153(7): p. 469-72.
41. Sox, H.C., et al., *Comparative effectiveness research: challenges for medical journals.* *J Clin Epidemiol*, 2010. 63(8): p. 862-4.

## Chapter 2. Research environment

### Overview

A foundational element of any research project is the research program environment. In the context of comparative effectiveness research using linked data, a secure and well-performing environment is important for several reasons, including that it helps build and assure trust between researchers and the providers of sensitive data – be it patients, registries administrators, insurance claims administrators, or others. If data providers are confident that a research partner has strong administrative and technical security systems and takes data security seriously at a programmatic level, they will be more confident in providing sensitive data to the researchers, including data with unique identifiers. As we describe in Chapters 4 and 5 of this report, linkage quality is typically much stronger when unique identifiers are available. Therefore, a secure research environment and capable technical information technology support can directly influence the quality of the research data obtained and, by extension, research results. With faith in the integrity and security of the research environment, data providers may also be more likely to provide other unique data that can be important to driving truly innovative research.

A secure and well-performing environment is also important in that system performance and security controls can directly influence the scope of the research project, including the size and complexity of the data that can be managed and linked to support that project. As the scope and complexity of research projects increase and the data volume grows, computing environments often are challenged to scale up to ensure seamless operations.

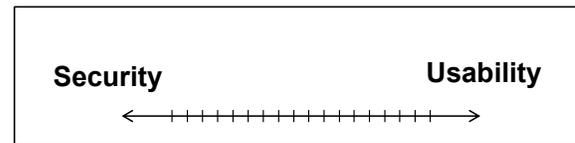
This chapter describes key research environment considerations, including the technical platform and security considerations, to guide researchers as they seek to develop or optimize their systems for CER projects using large volumes of data such as linked registry and administrative claims data.

### Computing systems and the balance between security and usability

As shown in Figure 1, security and usability often stand on opposite ends of a spectrum. The trade-off for having a highly secure system is decreased accessibility and practical usability, whereas systems that are highly accessible often face greater challenges in assuring data security. Understanding the scope of the research project and the needs of the researcher or research team are important to specifying a

system configuration that meets the needs of the project, and specifies and balances security and usability needs. For example, a single researcher with a small research project of limited scope will likely have different needs of a computing environment compared with a large, decentralized research team undertaking a multi-year research study using national data.

**Figure 1. Security versus usability**



We present three different computing system scenarios to help researchers identify where on the spectrum their program may fall. While this discussion does not take into account the number of users, it is important to note that cost of large systems vary greatly depending on existing infrastructures and purchasing prices of solutions offered by various vendors.

**Desktop computer: high security, low usability, low cost**

In this environment, the user accesses all data on a dedicated desktop computer located in a dedicated and locked office 24/7 with limited network access.

- *Security*: The risks of theft, network attacks and stolen equipment are reduced to a minimum.
- *Usability*: Multiple users will never be able to access the data concurrently.

**Central server: medium security, medium usability, moderate cost**

This environment allows multiple users to connect remotely through a secure command line (SSH) to a central computing server housing all the data and tools.

- *Security*: The risk increases while gaining access to information over a network. Controlling on what user can see what information creates new administrative challenges.
- *Usability*: Multiple users can collaborate on a central system. Computing jobs can be submitted in the background and the progress can be checked from remote locations.

**Virtual remote desktops: high security, high usability, high cost**

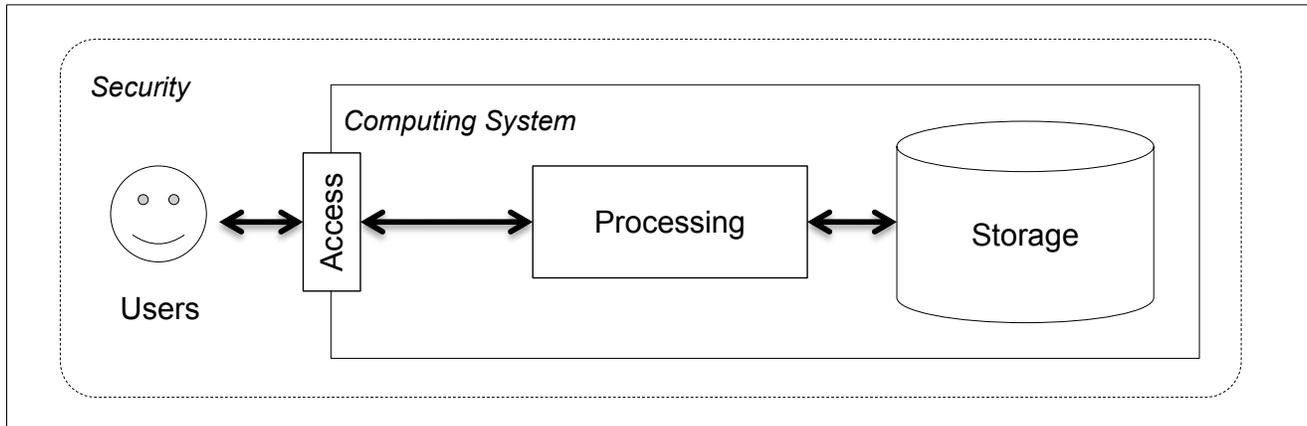
This environment allows multiple users to connect remotely from laptops or desktop computers to virtualized desktops over the internet to access shared data and tools.

- *Security*: Since the accessing computers only supply monitor, keyboard, and mouse, no data ever leaves the server environment. Even secure printing to dedicated printers is possible to control for paper output.
- *Usability*: Each user connects to a virtual computer in the central environment. All tools are presented in existing desktops.

## Building the technical platform

Regardless of the selected technologies, the number of users, or security requirements, the technical platform can be disassembled into various components. Defining the requirements for the individual components creates a reasonable descriptive plan.

**Figure 2: Component Architecture**



## Securing the platform

Most regulatory security frameworks such as the Health Information Portability and Accountability Act (HIPAA) and Federal Information Security Management Act (FISMA) focus on controlling the confidentiality, integrity, and availability of information. The efforts of implementing administrative, physical, and technical safeguards tend to scale up as the system complexity increases. Regulatory requirements and risk assessments will highly affect the technical implementations.

## Users accessing platform

To support a wide range of innovative research on novel data, the experience and focus of team members narrows and deepens. Work is divided between individuals to cover areas such as data management, data linking, cohort discovery, advanced modeling, and more. Complex research projects depend on the seamless integration and collaboration of the various users and the use of their preferred tools. By defining the user profiles and job responsibilities, the main usability properties of the expected environment are established. Examples of typical roles within complex project teams include:

### ***Role: Data Manager***

The data manager takes care of the data. This might include import of new data, conversion of file formats, preparation and receipt of data carriers, archiving of obsolete data, and granting access to data.

***Role: Data Linking***

The linking expert is responsible for linking data sources. This might include cleaning of linking variables, building linking methods, and cohort discovery resulting in datasets for various research projects.

***Role: Analyst / Statistician***

The statistician is responsible for all modeling aspects of a research study. This might include creating analytic cohorts for study questions (using previously linked de-identified data), preparing data for modeling, and analyzing data to meet project objectives.

**Gaining access to platform**

Access management controls how users gain access to the system and data. An existing organization might have a central user management process that establishes authentication with a simple username/password combination. More advanced two-factor authentication methods ensure that username/passwords cannot be shared, or in the case of biometric authentication (e.g., fingerprint reader) enforce the identity of the intended user. Examples of commonly used authentication methods and their pros and cons are provided in Appendix 1.

**Processing power of platform**

The processing power of the computing system directly affects the time it takes to manipulate the data. As linking processes touch the same information repeatedly, tuning and optimizing the performance of hardware and software parameters will reduce the run times. For considerations for hardware performance, see Appendix 1.

***Impact of network on data flow***

The network can quickly become the bottleneck for moving data, resulting in an exceptionally slow response. In an optimized setup, the connection between data and processing components is a dedicated Gigabit (1000Mbps) network or even fiber optics. Any components between the processing and storage such as firewalls, network switches, or routers will reduce information flow. In a setup in which the data are stored on hard drives directly attached to the processing system, network performance will have a limited impact on data flow.

**Data storage platform**

Because working with novel data requires exploratory processing of many files, the storage performance directly affects the time it takes to perform these tasks. The main technical characteristics of the storage

platform are size and speed. A storage device is attached using a specific technology such as Serial Advance Technology Attachment (SATA), Serial Attached Small Computer System Interface (Serial SCSI or SAS), Universal Serial Bus (USB), or Storage Area Network (SAN). In addition, various vendors in the market sell enterprise storage solutions encapsulating storage into one appliance.

### ***Storage size***

When purchasing data carriers it is important to understand that physical data size and actual available data size will greatly vary depending on the installation. Methods used to improve availability such as Redundant Array of Independent Disks (RAID) might require as much as twice the amount of physical space. The second limiting factor is the file system used to store data. A data carrier is divided into blocks like a blank book with many pages. The size of the data block is fixed for the entire file system. As an example: If the block size is 1024 characters (or bytes) and a file of 1500 characters is saved, it will consume 2048 physical bytes on the data carrier. Since the partially used blocks cannot be used for other files, these bytes are “lost”.

### ***Storage speed***

Storage speed has two properties: 1) the time it takes to find the data on the carrier (referred to as Seek Time in milliseconds), and 2) the continuous read/write performance. The Seek Time is mainly dependent on how fast the disk is spinning in Rotations per Minute (RPM). General values range from 5400, 7200, 10'000 to 15'000 and in case of a Solid State Disk (SSD), the seek time will be extremely low as there are no moving parts. The continuous read/write operations are dependent not only on how fast the disk is spinning, but also on how the disk is attached to the processing system.

A well-performing storage system can read/write information at rates of 100MB/s or more. A disk spinning at 5400 or 7200 RPM as delivered in standard laptops or desktops can generally not achieve this. In comparison, a Solid State Disk (SSD) attached over SATA can easily reach read/write rates of 500MB/s or more. To optimize cost, a computing system can be outfitted with slower/cheaper storage for archiving in combination with fast analytic storage to support powerful processing.

## **Securing the research environment**

### **Overview**

Federal and state level laws mandated regulatory requirements intended to control for one or more of the following objectives:

Confidentiality: Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information

Integrity: Guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity

Availability: Ensuring reliable and timely access to information

### **Regulatory requirements**

The research environment has to be compliant with applicable laws to protect the hosted information. Because HIPAA governs health information collected by covered entities mainly during health encounters, alternate research datasets might require compliance with other contractual requirements or local state regulations. Researchers should consult with a regulatory expert as early as possible to ensure that they understand the scope of all applicable laws.

Regulatory requirements generally describe WHAT must be controlled and leave it up to the research team to define HOW to reach the required controls by implementing adequate policies and procedures. Some state-level privacy laws may govern even self-collected information. In the following sections, we describe a sample of regulatory requirements that might be applicable to researchers working with sensitive health information.

#### ***Federal Information Security Management Act of 2002***

FISMA defines a mandatory framework for managing information security for all information systems used or operated by a U.S. federal government agency or by a contractor or other organization on behalf of a federal agency. It requires the development, documentation, and implementation of an information security program. The National Institute for Standards and Technology (NIST) standards and guidelines (Special Publications 800-series) and Federal Information Processing Standards (FIPS) publications further define the framework of the program.

#### ***The Health Information Portability and Accountability Act***

The US Congress established HIPAA in 1996. Security Standards establishing requirements to safeguard Protected Health Information (PHI) – both paper and electronic (ePHI) – were issued as part of HIPAA in April 2003. The security requirements specifically address administrative, physical, and technical safeguards meant to ensure that patient health records and personally identifiable information remain as secure as possible.

### ***State security breach laws***

Forty-six states, the District of Columbia, and multiple US territories (Guam, Puerto Rico, and the Virgin Islands) have enacted privacy-breach notification laws. While these laws can vary from state to state, they generally follow a similar framework. This framework includes a definition of “sensitive data,” requirements for triggering the breach notification process, identification of actors and roles in the notification process, defines to whom the law applies, and describes those cases under which certain parties and/or information may be exempt from this legislation.<sup>1</sup> Researchers are responsible for understanding their responsibilities under their relevant state breach notification legislation, and should consult legislative resources such as the National Conference of State Legislatures for regulatory text.<sup>2</sup>

### **Identifying sensitive data**

Sensitive data is the information protected by regulatory requirements. The definition of sensitive data varies widely between laws. In some cases, the scope of a “Data Use Agreement” (DUA) could even include restrictions on aggregated levels or define minimum cell sizes. In the following section, we provide a summary of regulatory definitions per FISMA, HIPAA and state security-breach laws.

### ***Personally Identifiable Information and FISMA***

As used in information security, Personally Identifiable Information (PII) is any information about an individual maintained by an agency. This includes 1) any information (e.g., name, social security number, date and place of birth, mother’s maiden name, or biometric records) that can be used to distinguish or trace an individual’s identity, and 2) any other information (e.g., medical, educational, financial, and employment information) that is linked or linkable to an individual. Examples of PII include, but are not limited to:

- Name, such as full name, maiden name, mother’s maiden name, or alias
- Personal identification number, such as social security number (SSN), passport number, driver’s license number, taxpayer identification number, or financial account or credit card number
- Address information, such as street address or email address
- Personal characteristics, including photographic image (especially of face or other identifying characteristic), fingerprints, handwriting, or other biometric data (e.g., retina scan, voice signature, facial geometry)

---

<sup>1</sup> <http://www.fas.org/sgp/crs/misc/R42475.pdf>

<sup>2</sup> <http://www.ncsl.org>

### ***Protected Health Information (PHI); HIPAA***

The HIPAA Privacy Rule protects all "*individually identifiable health information*" held or transmitted by a covered entity or its business associate, in any form or media, whether electronic, paper, or oral. The Privacy Rule calls this information "protected health information" or PHI.

"*Individually identifiable health information*" is information, including demographic data, that relates to the any following:

- the individual's past, present, or future physical or mental health or condition,
- the provision of health care to the individual, or
- the past, present, or future payment for the provision of health care to the individual, and
- information that identifies the individual, or for which there is a reasonable basis to believe it can be used to identify the individual.

Individually identifiable health information includes many common identifiers (e.g., name, address, birth date, Social Security Number). The Privacy Rule excludes from protected health information employment records that a covered entity maintains in its capacity as an employer, and education and certain other records subject to, or defined in, the Family Educational Rights and Privacy Act, 20 U.S.C. §1232g.

### ***Electronic Protected Health Information***

The HIPAA Security Rule protects a subset of information covered by the Privacy Rule, which is all individually identifiable health information a covered entity creates, receives, maintains, or transmits in electronic form. The Security Rule calls this information "electronic protected health information" (ePHI). The Security Rule does not apply to PHI transmitted orally or in writing.

### ***Limited Datasets; HIPAA***

HIPAA also has a provision for Limited Datasets (LDS) from which most but not all potentially identifying information has been removed. Elements in an LDS are often necessary for research; however, "Direct Identifiers" a subset of PHI defined by HIPAA §164.514(e) (2) must be removed. The direct identifiers include:

- Name
- Postal address information, other than town or city, State, and zip codes
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers

- Medical record numbers
  - Health plan beneficiary numbers
  - Account numbers
  - Certificate/license numbers
  - Vehicle identifiers and serial numbers, including license plate numbers
  - Device identifiers and serial numbers
  - Web Universal Resource Locators (URLs)
  - Internet Protocol (IP) address numbers
  - Biometric identifiers, including finger and voice prints
  - Full face photographic images and any comparable images
- Limited Datasets can include the following PHI:
- Dates of birth
  - Dates of death
  - Dates of service
  - Town or city
  - State
  - Zip code

***Personal Information; state security breach laws***

Researchers should review applicable state legislation for definitions of Personal Information.

Generally, these definitions do not vary substantially from state to state and are very similar to federal definitions. For example, the North Carolina State Security Breach Laws (North Carolina General Statute § 75-65) define Personal Information as a person's first name or first initial and last name in combination with any of the following identifying information:

- Social security number or employer taxpayer identification numbers
- Driver's license, State identification card, or passport numbers
- Checking account numbers
- Savings account numbers
- Credit card numbers
- Debit card numbers
- Personal Identification (PIN) Code
- Electronic identification numbers, electronic mail names or addresses, Internet account numbers, or Internet identification names
- Digital signatures
- Any other numbers or information that can be used to access a person's financial resources
- Biometric data
- Fingerprints
- Passwords
- Parent's legal surname before marriage

**Summary of protected information**

The research team may find it useful to summarize the protected information types identified by their applicable regulatory requirements. This matrix will help the research team identify information in

datasets and assess the policies and procedures that might apply to a specific work task. Table 1 shows an example of one such matrix.

**Table 1. Example of matrix summarizing protected information types**

Identifying Information	Sensitive Data Type				
	PHI	ePHI	PII	Personal Information	Direct Identifiers
Name (Full name / maiden name / mother's maiden name / alias)	X	X	X	X	X
Address Information	X	X	X		X
Telephone/Fax Information	X	X			X
Personal IDs (SSN / Taxpayer ID / Driver's license number / State ID / Passport number / Birthdate / Certification or License numbers)	X	X	X	X	X
Financial IDs (Checking/Savings account numbers / PINs / Credit Card numbers)	X	X	X	X	X
Electronic IDs (Email name/address, internet account numbers, internet ID, passwords)	X	X	X	X	X
Personal Characteristics (Digital Signatures / Biometric data / fingerprints / handwriting / full face images)	X	X	X	X	X
Healthcare data / provisions / payment / beneficiary information (Past, present, or future)	X	X			X
Employment Information	X	X			X
Device IDs / Serial Numbers	X	X			X
Vehicle IDs	X	X			X

### Building and implementing a security plan

Meeting applicable regulatory requirements requires thoughtful planning and management. While it is tempting to think of information security in terms of technological controls, successful security management requires people, processes, and technology in equal proportion. An overarching security management plan addresses how people, processes, and technology will be leveraged to maintain the confidentiality, integrity, and availability of sensitive data within the bounds set by applicable regulatory requirements.

While development of the security management plan is an iterative process, with sections added or refined as planning activities proceed, the document will ultimately address the following:

- *Security Laws and Regulations* describes those regulatory requirements applicable to the research team as discussed previously
- *Major Functions* lists those functions the Security Plan is intended to accomplish
- *Scope* lists those sensitive data types the security program is intended to address
- *Roles and Responsibilities* describes roles which will be held by members of the organization and their responsibilities vis-à-vis information security

- *Management Commitment* represents an official statement on the part of the applicable management body in support of the processes and procedures documented within the security plan
- *FISMA Security Categorization and Impact Level* defines the FISMA category assigned to the data and information systems covered by the security plan. Please note that this section is only applicable to those systems subject to FISMA.
- *Compliance and Entity Coordination* describes which role(s) is responsible for ensuring organizational compliance with the security plan, and which role(s) is responsible for coordinating security activities among relevant entities external to the research team (e.g., data centers, overarching security offices, etc.)
- *Implementation and Governing Plans* describe, at a high level, the number and content of all security sub-plans defining the processes and procedures for:
  - Security Documentation Control
  - Risk Management (described in further detail below)
  - Workforce Security
  - Access Management
  - Security Training
  - Incident Reporting
  - Contingency Planning
  - Security Assessment
  - Facility Access
  - Workstation Access
  - Device and Removable Media
  - Data Integrity
  - Authentication
  - Network Security
  - System Activity Review/Audit

At the outset of security planning, the research team should be able to define the Security Laws and Regulations, Major Functions, and Scope sections. Roles and Responsibilities, Management Commitment, Entity Coordination, and FISMA Categorization (if applicable) can be defined further through stakeholder meetings. The processes and procedures documented in the sub-plans will be developed as part of the risk management process described below.

### **Workforce Training**

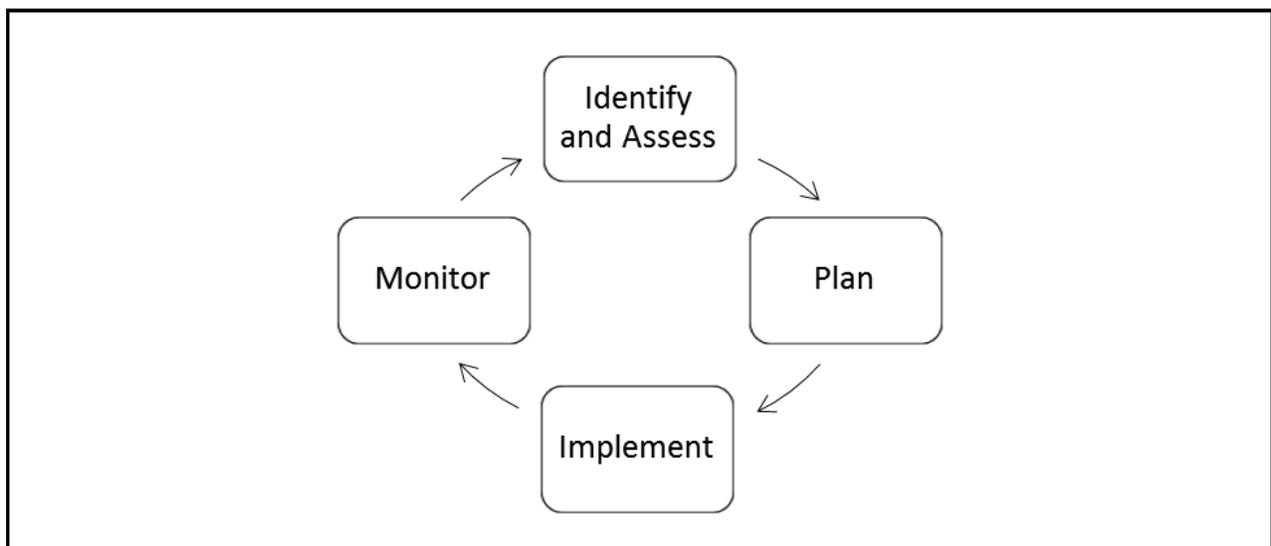
A training plan defines working procedures, emergency and incident management, sanction policies, policies and procedures on how to inform members of the workforce about their roles and responsibilities, and other relevant procedures. Many larger research environments might be able to leverage existing training modules as a baseline. These might include training on HIPAA, research ethics, basic computer and network use, and basic HR policies. Keeping the re-training on an annual basis is advisable.

## Managing risks

Research teams must first understand regulatory requirements, then select and implement adequate security controls to meet these requirements and to mitigate other potential risks posed to the security of the organization's information systems and data. Often, discussions of information security jump directly into discussions of specific technical safeguards, mistakenly emphasizing the importance of technical solutions rather than the risks these solutions are designed to control. An emphasis on risk management, however, properly defines technical solutions as the means by which organizational risks are controlled. Risk management, therefore, drives information security planning. A comprehensive risk management program not only allows data custodians to identify risks posed to their data, but also provides a framework for selection of functional and technical security controls.

Data custodians subject to FISMA requirements should consult the National Institute of Standards and Technology's (NIST) guidance for implementing a FISMA-compliant life cycle program, which includes detailed volumes of guidance and controls. We illustrate a more general risk management framework in Figure 3. This framework envisions risk management as a continuous cycle of assessing, addressing, and monitoring organizational risk to ensure the confidentiality, integrity, and availability of information systems.

**Figure 3. Risk Management Cycle**



## Identifying and assessing risk

Risk identification is conducted on any technology, process, and procedure within the scope of the environment. Risk identification is, simply put, the process of identifying and documenting potential threats to the research team's information and information systems. Risk identification can be conducted in a variety of ways, including brainstorming sessions, documentation reviews, assumptions analysis, cause and effect diagramming, Strengths/Weaknesses/Opportunities/Threats (SWOT) analysis, and expert consultation. Inclusion of an independent third party, be it an outside consultant or even representatives from a separate group within the research team, will provide an external point of view invaluable in fully mining the spectrum of potential adverse events. Regardless of the method used, this process must clearly identify and document the source of the risk and the impact of the risk should it be realized.

Assess identified risks along two primary dimensions: 1) probability of occurrence, and 2) criticality of impact. Actions and mitigations planned in the next phase of the risk management cycle will be based largely on each risk's score as assessed during this phase.

### Scoring risk

Risk scores evaluate the combination of the probability and impact of a security breach/incident. Higher scores represent higher security risks. Lower scores represent reduced security risks. Table 2 is an example of risk scores.

### Tracking risks

The risk register is the collection of all identified risks, their assessed impact and probability, and possible actions/mitigations. Both HIPAA and FISMA mandate the analysis of risk and a record thereof. Table 3 on the next page is an example of a generic risk register.

**Table 2. Example of risk scores**

		Criticality of Impact		
		Low	Medium	High
Probability of Occurrence	High	3	6	9
	Medium	2	4	6
	Low	1	2	3

Examples of Impact:

- Loss of data
- Loss of public confidence
- Potential lawsuit
- Multiplier effect

Examples of Occurrence:

- Frequency of information access
- Number of users
- Size of dataset

**Table 3. Example of generic risk register**

#	Score	Title	Description	Impact	Impact Score	Probability Score	Action
R001	6	Equipment stolen from office site	An individual gains access to office and removes electronic equipment.	Leak of data; through loss of computers, access to secure systems might be compromised	High	Medium	Lock building, encrypt data, alarm system
R002	3	Network Access	An individual gains access to the network within the organization.	Leak of data, stolen passwords; computers connected to network are subject to hack or sniff attacks.	High	Low	Network firewalls, internal switched network, virus software, secure wireless
R003	6	Remote Support	Support gains access to an organizational computer for remote support purposes.	Leak of data; open connections to protected systems and open windows displaying PHI are visible to support individual.	Med	High	Configure to no auto desktop sharing, close applications with PHI
R004	6	Computer Crash or Upgrade	A computer used to access PHI stops working or is upgraded.	Leak of data, leak of passwords; old hard drive might get re-used, warranty traded or sold	High	Med	Disk encryption; disposition policy; no saving of PWDs, use password mgmt software
R005	2	Lack of DUA understanding	DUA limitations might be misinterpreted.	Violation of contract, misuse of data, leak of data	Low	Med	Governance DUA, training
R006	2	Unauthorized staff gains access to data	Staff, students, collaborators, reviewers gain access to data protected by DUA.	Violation of contract, misuse of data, leak of data	Low	Med	Administrative safeguards, DUA
R007	6	Data emailed for review	Protected data is shared through email, electronic documents within the intent of review.	Leak of data, violation of contract; information is no longer housed on secure environments.	Med	High	Secure email, training of email use, secure drop box
R008	2	Data emailed to incorrect individual	An email with protected information is sent to a recipient it was not intended to.	Leak of data	Low	Med	Email training, secure drop box

## Planning risk responses

Once risks have been identified, assessed, and documented in the risk register, data custodians and other stakeholders can plan appropriate methods of dealing with each risk. Risk responses can be divided into one of four categories: avoidance, acceptance, transfer, or mitigation.

- *Risk avoidance* occurs when a research team takes the necessary actions to reduce the likelihood of risk realization to close to (if not exactly) zero. Generally, risk avoidance is the most desirable method of dealing with risks; however, it is often cost prohibitive or simply infeasible to avoid all risks.
- *Risk acceptance* occurs when a research team chooses to accept the consequences of a risk should it be realized. Risk acceptance is generally recommended when the impact of a risk is small or when the probability of occurrence is significantly lower than the cost to avoid, transfer, or mitigate. Each research team must define its own criteria for what constitutes an “acceptable” risk.
- *Risk transfer* occurs when a research team passes the impact of risk realization on to another party. The most commonly experienced form of risk transfer is insurance. Risk transfer is only feasible when the impact can be clearly measured and addressed by the third party.
- *Risk mitigation* occurs when a research team takes steps to reduce the probability or impact of risk realization. Risks that cannot be avoided, accepted, or transferred must be mitigated. After a research team decides whether to avoid, accept, transfer, or mitigate a risk, it must determine the necessary steps to do so. At this juncture, the research team will identify the necessary and appropriate technical controls to either avoid or mitigate certain risks. Actions identified in this phase must also be documented in the risk register.

## Implementing risk responses

During the implementation phase, the research team develops and deploys the technical controls identified in the planning phase. Just as critically, the research team must also document the controls selected, develop all necessary records, and train stakeholders accordingly. Users must understand not only how to use any security controls implemented but also the “rules of behavior” for maintaining a secure environment. Technical controls alone are not sufficient to create a fully secure environment; users and other stakeholders must foster and maintain a culture of security.

## Monitoring risks

Risk management is an ongoing cyclical process. The research team must periodically re-assess the environment for new or changing risks, which in turn must be identified, assessed, addressed, and their responses implemented. Thoughtful and frequent monitoring of risk allows a research team to more easily adapt to changes, both expected and unexpected, without compromising information security.

## **Conclusions**

A research program environment incorporating a secure and well-performing computing platform represents the operational backbone for conducting innovative research using novel data. Securing and safeguarding the information is not only required by law through regulatory frameworks, but on successful execution it builds the needed trust among stakeholders. Data providers will be more open to providing access to their information, researchers will be confident in accessing the resource even with sensitive data in use, and programmers and analysts become efficient through the application of technologies and tools. The technical implementation combining performance and storage will enable the complex data management as described in Chapter 3. Supported by the leadership, the successfully implemented security policies and procedures seamlessly fit into daily workflows reducing and mitigating potential risks. The environment is now ready to conduct research and receive any type of data regardless of its sensitivity.

## Appendix 1. Procedures and processes to enhance data security

### Moving sensitive data using CDs or DVDs

Perform the following steps to create and transport sensitive data using CDs or DVDs. This procedure can also be adapted to the use of transferring data over electronic file transfer such as Secure File Transfer Protocol (SFTP).

- 1) A new media number is created and added to a data carrier list tracking all moveable media containing sensitive data.
- 2) Create a local folder with the media number as the name.
- 3) Assemble all the data in the created folder.
- 4) Generate an encryption key using a GUID<sup>3</sup> tool such as <http://www.guidgenerator.com/> and print it on a document along with the media number.
- 5) Create an archive using a PGP<sup>4</sup> encryption tool with all the contents of the folder using the GUID as the encryption key.
- 6) After testing the self-extracting archive, use a file shredding tool<sup>5</sup> to remove the folder with the data.
- 7) Burn the archives onto a data CD or DVD labeled with the media number.
- 8) You can now mail the data carrier and fax or email the encryption key separately to the receiving party.

### Guidelines for storage and destruction of moveable media

- Store moveable media in a safe, separated from the encryption keys.
- Destroy damaged and/or retired media, including hard disks, by shredding.
- Update moveable media records for every media item that is disposed or destroyed.
- Shred any printed material at location or use a secure document disposal service.

### De-identification and mapping of data

The process shown in Figure 4 describes how to add data sources to the research data environment, and how to remove direct identifiers and create a merged dataset with elements regarding the individuals' health status or health services utilization.

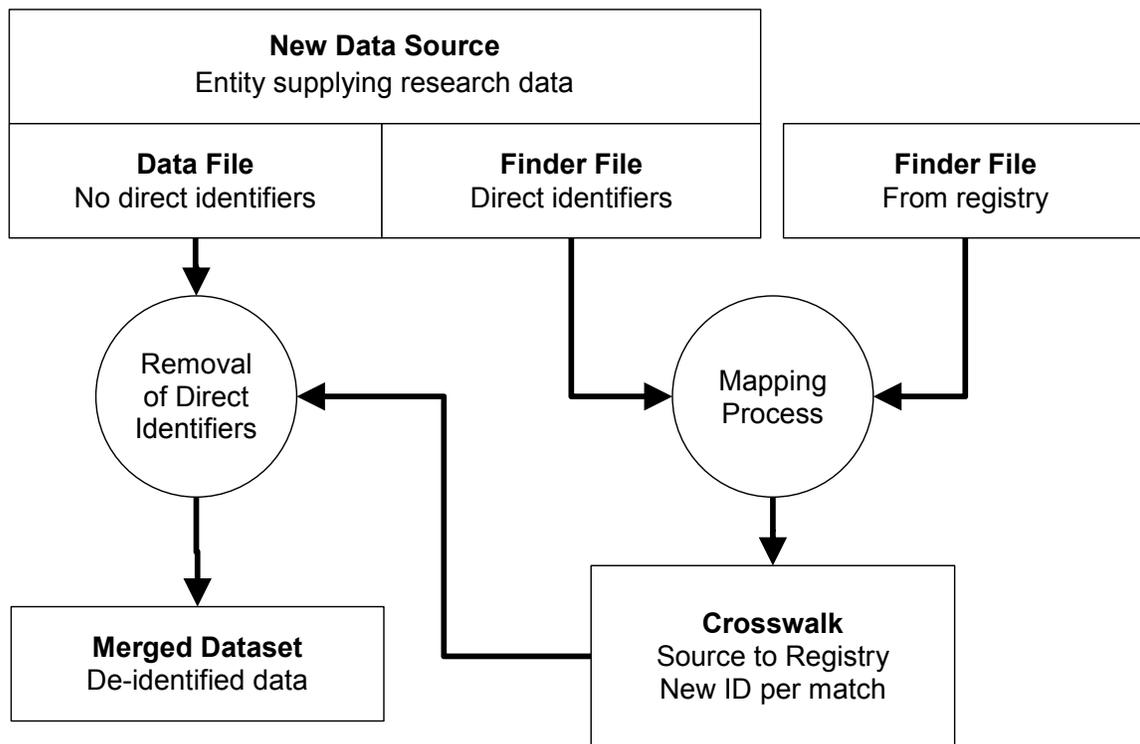
---

<sup>3</sup> Globally Unique Identifier

<sup>4</sup> [http://en.wikipedia.org/wiki/Pretty\\_Good\\_Privacy](http://en.wikipedia.org/wiki/Pretty_Good_Privacy)

<sup>5</sup> <http://www.fileshreder.org/>

**Figure 4. Process for creating a merged dataset**



Data File: This file contains raw data such as claims, diagnoses, treatment, etc. Individuals cannot be directly identified in these data.

Finder File: The finder files contain Direct Identifiers of individuals.

Mapping Process: Using a mapping method, the individuals in the finder files are matched

Crosswalk: The crosswalk identifies the same individuals and possible duplicates. A new unique identifier (ID) is assigned per true identified research subject.

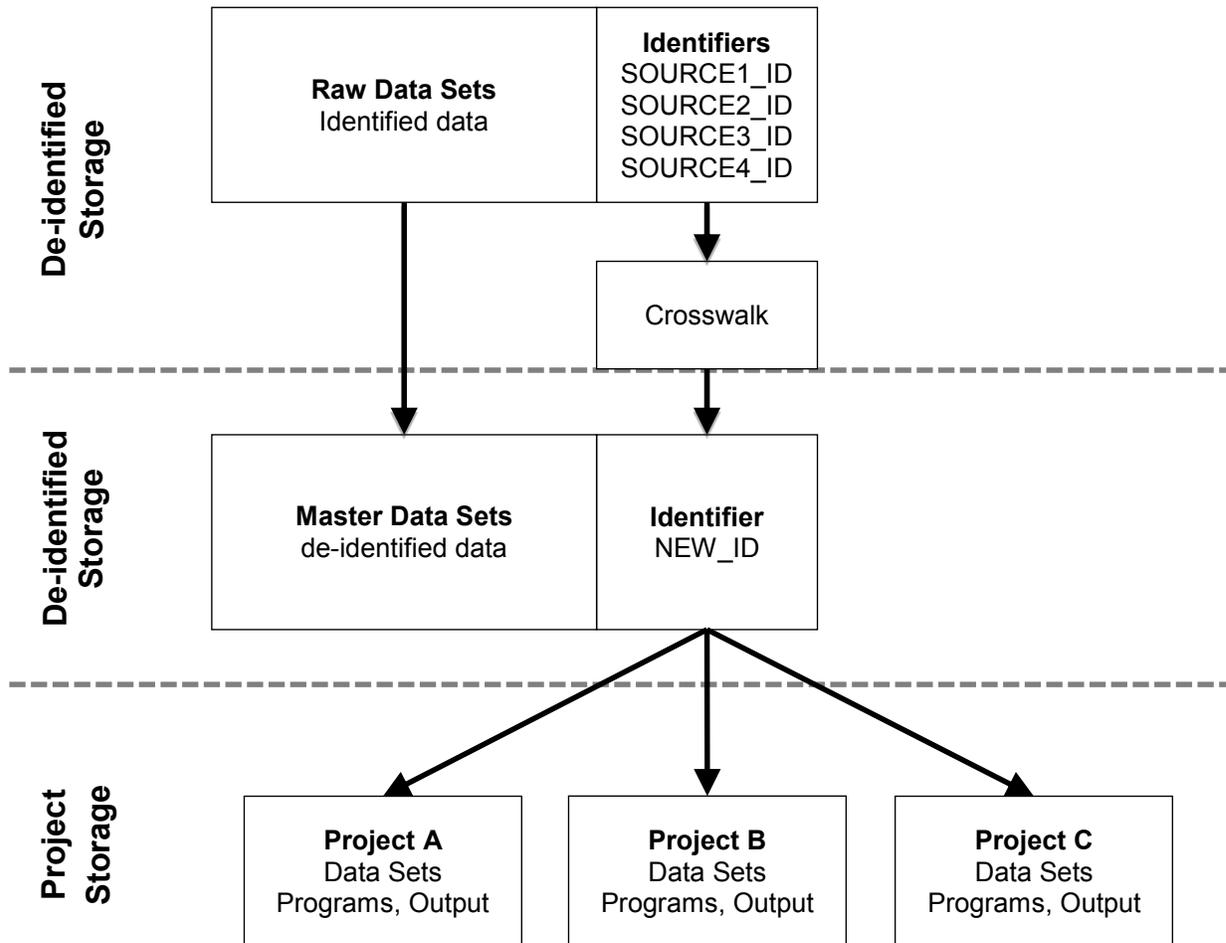
Removal of Direct Identifiers: Using the “Crosswalk” and the “Data File” a new file is generated by replacing the source identifiers with the new IDs.

Merged Dataset: Information about an individual can now be access across data sources and de-identifies datasets using the new IDs.

## Physical separation of data using storage architecture

In the computing environment, data are separated physically into the three storage areas as shown in Figure 5.

**Figure 5. Physical storage areas**



Direct Identifier Storage: Raw data including direct identifiers is secured in a highly restricted part of the system. The “Direct Identifier” space is accessible only through very restrictive access management controls. All work on data linking and removal of direct identifiers is performed in this environment. Only a limited set of specifically trained, authorized individuals work in this environment.

De-identified Storage: The “De-identified” space contains de-identified “Master Datasets” from each source. This storage might be accessible as read-only only to authorized users. Research datasets are extracted or “cut” from these master files.

Project Storage: Individual programmers access this space for projects. IRBs and DUAs control the access to master datasets and datasets for projects.

## Examples of common authentication methods

### Simple user account

A user authenticates with username and password.

- *Pros:* This is a quick and simple way to access a system.
- *Cons:* Once the username and password are known, any individual gains access.

### Two-factor authentication

A user authenticates with username, password, and PIN (e.g., RSA Secure ID)

- *Pros:* A changing PIN provides a second factor in addition to username and password. An authentication is not possible without the device.
- *Cons:* The PIN verification is costly and, depending on the implementation, requires the computing environment to have access to the system verifying the PIN.

### Biometric authentication

A user authenticates with username, password, and a fingerprint.

- *Pros:* The additional biometric verification requires the account holder to be present at time of authentication.
- *Cons:* Managing biometric information requires custom software installations on client system.

## Performance considerations for computer hardware

### Memory, Central Processing Unit, Bus Speed Parameters

Fine-tuning these three hardware components is essential for the actual processing power. Because Random Access Memory (RAM) can be added after a purchase, changes to the CPU or bus speeds are complicated and impractical. The Bus Speed represents how fast information can be moved between the CPU and the Memory, and from and to the attached storage. When purchasing a system, you should purchase the highest affordable CPU/Bus Speed combination while leaving space to add memory later.

#### *Central Processing Unit*

Performance of a Central Processing Unit (CPU) is affected by a) sockets (i.e., how many actual CPUs), b) cores and threads (i.e., how instructions can be processed in parallel), and c) clock (i.e., a number in GHz representing how many instructions are processed per second).

#### *Random Access Memory (RAM)*

Selecting memory is dependent on the supported architecture of the motherboard. Manufacturers generally advise on what type of memory is supported and best for optimized performance. The product description of memory often includes physical and performance parameters. For example “240-Pin

DDR3 1600” represents memory that has 240 pins connecting it to the motherboard while supporting a 1600MHz clock.

### ***Bus speed***

Depending on the architecture, there are multiple bus speeds affecting data throughput between storage, memory, and CPU. For example: a hard disk might be attached using SATA3 or “SATA 6Gb/s” representing a bus speed of 6Gb/s. Bus speeds on the CPU/motherboard (chipset layout) are more complicated and fine-tuned by the vendor. Components with fancy names such as North Bridge, South Bridge, FSB (Front Side Bus)<sup>6</sup> are part of this architecture. Vendors typically offer systems with high performance options where these speeds are optimized and enhanced compared with home use products.

---

<sup>6</sup> [http://en.wikipedia.org/wiki/Front-side\\_bus](http://en.wikipedia.org/wiki/Front-side_bus)

## Chapter 3. Linkage feasibility—to link, or not to link

### Overview

In this chapter, we provide an overview of factors that researchers should consider before embarking on a data linkage project. These include:

- 1) Determining the feasibility of data linkage through assessing variable overlap and data quality
- 2) Determining the original purpose of the data and whether the terms of the original data collection prohibit the linkage
- 3) Determining data ownership, regulatory requirements, and limitations on use
- 4) Planning for data sharing and managing data security concerns, and
- 5) Ensuring a qualified team is available to manage data security and the technical aspects of the data linkage.

Because of the importance of each of these steps, researchers should ideally evaluate these concerns to the extent possible before applying for grant funding (see *Chapter 6* for a checklist for preliminary planning steps).

### Evaluating linkage feasibility

A fundamental step in any linkage effort is the prospective assessment of linkage feasibility. The feasibility of a linkage project depends largely on the quantity and quality of the identifying information available in the data sources being linked (Newcombe, Smith et al. 1983).

Identifiers, full or partial, are more or less informative depending on their discriminatory power, or number of unique values. For instance, month of birth, which has 12 unique values, is more informative than sex, which has only two unique values. Assuming normal distributions, record pairs matched randomly will agree on sex 50% (1/2) of the time simply by chance, while record pairs matched randomly will agree on month of birth 8.3% (1/12) of the time. Thus, when a matched pair agrees on month of birth, it is less likely that the pair matched simply by chance, and more likely that the pair matched because the records represent the same individual. In this sense, month of birth contains more information than sex. Combining identifiers further increases the number of unique values (or “pockets”) thereby decreasing the chances that two records matched on the combination of identifiers matched by chance alone. Table 4 shows identifiers commonly used for record linkage.

**Table 4: Linkage identifiers**

Type of Identifier	Variable Description
Unique Patient Identifiers	Social Security Number Medical Record Number Patient / Beneficiary Identification Number
Indirect Identifiers	Name: first, last, middle, maiden, alias Dates: birth, death, diagnosis, treatment, admission, discharge Sex Geographic Location: street, city, county, zip, state Diagnosis Codes

Additional information can be gleaned from the values themselves. That is, matches on rarely occurring values compared with matches on frequently occurring values are less likely to occur by chance. For instance, a match on a rarely occurring surname such as “Lebowski” is less likely to occur by chance than a match on a frequently occurring surname such as “Smith.” A match on a rarely occurring surname therefore increases your confidence that a matched pair is a true match more so than a match on a frequently occurring surname.

Before embarking on a linkage project, it is important to consider whether a reliable and accurate linkage is possible given the available identifiers, and their discriminatory power. Potential data quality issues that cannot be known before receiving the data prevent us from knowing this for certain. For instance, every beneficiary in a claims database may have a valid SSN, but we cannot know ahead of time whether the SSN was entered correctly or whether it represents the given individual (e.g., in a claims database, the SSN for the primary subscriber may also be used for the subscriber’s spouse and/or children). However, using prior work in information theory, researchers have the means to estimate the discriminatory power of the available identifiers and the chances of uniquely identifying an individual within a single dataset or across multiple datasets.

The discriminatory power of a given identifier or set of identifiers can be quantified as the sum of  $abs(p \cdot \log_2(p))$ , where  $p$  is the proportion of records captured by each unique value of that identifier or set of identifiers. (Shannon 1948) Say we have a simple dataset of one variable (sex), and three records, one male and two females. In this scenario, the discriminatory power of the variable, sex, is equal to  $abs((0.33) \cdot (\log_2 \cdot 0.33)) + abs((0.67) \cdot (\log_2 \cdot 0.67)) = 0.92$ . Using this method, the discriminatory power of each available identifier or set of identifiers can be measured, and identifiers or sets of identifiers can

be ranked from most to least discriminatory. Notice that two variables with the same number of unique values can have varying levels of discriminatory power depending on the distribution of records across the unique values.

Additional work has led to the development of methods to assess individual identifiability or record uniqueness in a given dataset.(Roos and Wajda 1991; Howe, Lake et al. 2007) These methods examine the frequency distributions for every possible combination of variables in the dataset. Those variables with the highest number of unique categories and those variable combinations that identify records uniquely have the best chance of accurately identifying the entity that the information represents. For the purposes of record linkage, the minimum set of variables needed to successfully link two or more datasets is that combination of variables for which each record is identified uniquely (the mean number of records in each pocket is  $\sim 1.00$ ).(Roos and Wajda 1991) Using this method, researchers can request that each of the data vendors examine the percentage of unique records in their dataset identified by each combination of variables available for release. Variable combinations that approach the threshold of  $\sim 1.00$  records per pocket in both of the datasets to be linked are likely to succeed. The  $\sim 1.00$  threshold can be relaxed for researchers engaged in exploratory research. Researchers testing hypotheses are advised to adhere to the  $\sim 1.00$  threshold as false positives and false negatives will likely add bias to subsequent statistical analyses. SAS code for assessing record uniqueness in a dataset has been provided by Tiefu Shen and can be found at [www.naaccr.org/Research/DataAnalysisTools.aspx](http://www.naaccr.org/Research/DataAnalysisTools.aspx) under “Record Uniqueness.”

Cook and colleagues have developed a method to determine whether more sophisticated methods, such as the probabilistic techniques that will be discussed in *Chapter 4*, can be used to link the given data sources.(Cook, Olson et al. 2001) This method determines the level of discriminatory power needed to link two records with a certain desired degree of confidence (e.g., 95%), by comparing the combined discriminatory power (assessed as weights) over all available variables with the difference between the current weight and the desired weight. This method is discussed in detail in *Chapter 4*. Briefly, it allows researchers to set a threshold for the discriminatory power needed to successfully match files (an information threshold) while meeting a pre-specified false positive rate. Once this threshold is met, researchers should consider the minimum discriminatory power needed so as not to exhaust funds obtaining more information than is required, or unnecessarily compromise subject confidentiality.

Knowing how much discriminatory power is contained in each data source and the number of overlapping identifiers available allows the researcher to determine whether linkage is possible, and informs future decisions regarding the minimal set of identifiers required to assure a high quality linkage. However, establishing feasibility is not merely an issue of whether common variables exist, but how much overlap exists for the data sources in both content and temporality.

Regarding content overlap, two datasets may have similar variables but very different approaches for coding data or handling missing data that could influence the quality of the data linkage. For example, assume that one dataset codes missing variables as “.” and the second dataset codes missing variables using interpolated values. While it is possible to establish a mechanical record link in this setting, the linked data may not provide useful information or, worse, may be misleading. As another example, assume that researchers are linking health plan claims data and electronic medical record data from two sources in an attempt to capture more fully information on vaccine receipt. Assume also that each data source has a variable called “VACCINE.” If one source uses a “1” to indicate a vaccine was given and another source uses a date to indicate when the vaccine was given, researchers may find that the information they hoped to gain through the linkage is fundamentally flawed and cannot be used reliably. Thus, it is important for researchers to understand not only that a variable exists, but also the meta-data related to that variable, including how and why it was captured in that form.

Researchers often wish to link data sources that are refreshed, updated, or captured on different cycles. These situations with temporally changing data pose challenges due to dynamic changes in key linkage variables. For example, addresses, phone numbers, and names (e.g., women who get married) change over time. If there are gaps between when this information is collected for each data source, these variables may negatively affect linkage success. Additionally, temporal disparity between datasets would affect the functional data linkage utility, but may be undetected in the mechanical linkage processes. For example, if record sets are linked from sources having different collection timeframes, it may be possible to correctly link the databases mechanically, but the information from the sources will not be synchronized. This will cause misinterpretation of the joint data patterns present in the record merge. Conversely, the mechanical link can fail to establish identity, since the reference variables can change for any individual over fairly short time intervals. This is an important consideration for both the evaluation of variable overlap and ultimately for the appropriateness of inferences made using the linked data.

## **Purpose and conditions for original data collection**

With the evolution of data linkage techniques, researchers are now able to link multiple sources of data, including administrative health plan claims, disease registries, cohort studies, electronic medical records, financial data, and environmental data (and so on). In addition to confirming that data sources can be linked from a technical aspect, researchers must also consider ethical and legal restrictions that may limit the use of each data source that the investigator wishes to link. Two hypothetical scenarios are provided below.

*Scenario 1:* A researcher wishes to link state cancer registry data to administrative health plan claims data from a private health insurer to determine the impact of provider characteristics on the likelihood of a patient receiving a particular treatment. However, the insurer will not release provider information due to separate privacy agreements in place between the insurance company and physicians.

*Scenario 2:* A researcher wishes to link cohort data (previously collected in a prospective cohort study) to administrative health plan claims to evaluate the rate of sexually transmitted disease or mental health conditions for people who participated in the cohort study. As part of the original study agreements, patients in a cohort study were told that potentially sensitive health information (e.g., treatment for sexually transmitted disease, treatment for severe mental illness) would not be collected or used as part of the research,

These hypothetical scenarios are presented to demonstrate the importance of understanding the original purposes for the data sources that the researcher hopes to use for data linkage, including limitations or conditions for data use that might affect the proposed research. Investing the time to ensure that the research objectives can be met without negative consequences for involved parties is an important step in assessing the feasibility of data linkage.

## **Ownership of data**

In addition to practical and ethical considerations described above, determining ownership of data and restrictions on data use are necessary steps for researchers who want to link data. First, identifying the “owner” of each data source will ensure that the ethical and practical considerations noted above can be addressed in the research-planning phase. Next, the researcher can begin to outline a data governance process that will ensure the original data owners are informed of and approve the use of their data for the

study in question as well as any subsequent uses of the resulting linked dataset. Generally, the data governance plan should define:

- **Who** owns the data, who must grant approval for data use, and who may be granted access to the data;
- **What** regulatory requirements the data are subject to and what is considered “acceptable use” of the data (e.g., it is possible that the data, when linked, will be subject to additional regulatory requirements not pertinent to the original unlinked data);
- **Where** the data will reside;
- **When** the various stakeholders identified above are brought in to the governance process (e.g., are there requirements for original data owners to review and approve manuscripts before submission for publication with a peer-reviewed journal);
- **How** data-use approvals will be granted and how the data will be managed and secured; and
- **Why** the data are to be linked—what is the overarching purpose behind and goal for this linkage.

Proactive development of a data governance process provides several benefits to the researcher. First, it will help identify potential limits on publication and use of the data before study initiation. Data vendors with little experience working with academic researchers may require more information regarding the publication process before defining data use limits. Additionally, governance definition will inform the security infrastructure development and risk management planning discussed in *Chapter 2* by identifying who will have access to the data, where data will reside, and how data will be managed. Finally, a well-defined and consistently enforced data governance process will demonstrate to potential data owners that the researcher is committed to acting as a “good steward” of the data and building a relationship as a trusted partner to the data vendor.

The data governance process will both inform and be informed by the nature of the data-use agreements (DUAs) between the researcher and data partner, and the regulatory requirements for the respective data. Proactive planning when developing the DUA will allow the researcher and data partner to identify and discuss the parameters for all potential uses of the data, including whether to allow for future linking to additional datasets, whether the data partner will require review and approval of resulting research products (e.g., manuscripts, abstracts), and when. Particularly important for academic researchers, these agreements may limit publication of findings; it is important that negotiations take

place early in the data governance process to ensure that dissemination of evidence is not limited and that expectations of the original data owner and the researcher are clearly outlined.

The governance process should likewise document the nature and extent of the required Institutional Review Board (IRB) review. Large multi-site studies will require a higher level of coordination among institutional review boards. Researchers with such studies are strongly advised to consider centralized or federated IRB structures. The vast majority of individual linking projects, however, will require review only by their institutional IRB. Researchers intending to produce a linked dataset for use in multiple studies, whether single or multi-site, should strongly consider developing and seeking IRB approval of an “umbrella” protocol. This protocol would “cover the establishment of the research database, including information on items like data processing and governance” (Marsolo, 2012) The umbrella protocol can be amended to incorporate subsequent studies, if necessary, or subsequent studies can submit separate protocols referencing the umbrella protocol to the IRB.

## **Data sharing and security concerns**

As mentioned above, a well-defined and consistently enforced data governance process demonstrates the researcher’s desire to be a trusted partner to data vendors and others who manage large datasets. Critical to establishing and maintaining this relationship are documented processes and procedures for securing identifiable information.

Adequate planning to protect identifiable information is crucial to receiving authority to move forward with studies which otherwise may not be allowed. We provide guidance for researchers who are building their own security systems in *Chapter 2*. However, not all researchers have the resources available to build such systems. In addition, not all data partners/vendors are comfortable with releasing private health information or patient-level identifiers to researchers. There are several options available to researchers who still wish to link data sources but who encounter resource limitations or who are unable to negotiate successfully for data access from a partner. In such cases, researchers should work with data partners/vendors to determine whether hash encryption, third party linkage, and honest brokers would satisfy security needs to allow the project to move forward.

When existing identifiers may not be released or transmitted, records can be linked effectively using identifiers encrypted before original data holder release using cryptographic hash functions (Quantin,

Bouzelat, et. al., 1998; Quantin, Bouzelat, et. al., 1998(2); Quantin, Bouzelat, et. al., 1998(3)). An example is the National Security Agency's Secure Hash Algorithm v2 (SHA-2), published by the National Institute of Standards and Technology (Schneier B., 1994). SHA-2 values cannot be reverse engineered, and are widely used in security applications. Applications typically include formatting the input and concatenating a prefix, one or more variables, and a postfix. Formatting these data elements identically and applying the SHA-2 algorithm using the same key will generate the same encrypted identifier in each dataset, which can be linked deterministically to other datasets that use the same formatting, data elements, and key. Actual identifiers may be stripped and the datasets still merged, allowing data partners to limit any risks related to transferring data with PHI for research purposes.

Another option is to involve a third-party vendor or honest brokers to act as an independent source for managing data. Third-party vendors typically have extensive experience managing PHI and securing data, which can be reassuring for vendors who are wary of transferring data to researchers. Honest brokers provide a "firewall" between clinical or private health data and researchers. Individuals identified as "honest brokers" or third party vendors are generally selected to ensure the privacy of data and to minimize any conflicts of interest that could exist (Dhir, R., Patel A. et.al, 2008).

### **Building the Team**

A "team science" approach is required to successfully develop and execute a linking project. This involves multi-disciplinary collaborations in which experts from different fields share insights, perspectives, and tools from their respective disciplines. These individuals need to be technical experts but also must be able to bridge technical and disciplinary gaps and communicate effectively across the team to achieve a cohesive solution (Gladwell, M, 2000). However, true "team science" must be "transdisciplinary", defined as the integration of these disciplines together within the research framework which moves the project beyond what would be achieved by any of the individual disciplines alone (Stokols, D., Hall, K. et. al., 2008; Hall, K., Feng, A. et.al, 2008). Linking data effectively requires identifying individuals from diverse fields including population science, information science and technology, [bio]statistics/analytics, and regulatory/legal project management. It is important to identify and recruit potential team members in the early stages of the linkage project, because each will have important and specific insight regarding feasibility.

## References

1. Newcombe HB, Smith ME, Howe GR, Mingay J, Strugnell A, Abbatt JD. Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Comput Biol Med.* 1983;13(3):157-169.
2. Shannon C. A Mathematical Theory of Communication. *The Bell System Technical Journal.* 1948;27:379-423.
3. Howe HL, Lake AJ, Shen T. Method to assess identifiability in electronic data files. *Am J Epidemiol.* Mar 1 2007;165(5):597-601.
4. Roos LL, Wajda A. Record linkage strategies. Part 1: estimating information and evaluating approaches. *Methods of Information in Medicine.* 1991;30:117-123.
5. Cook LJ, Olson LM, Dean JM. Probabilistic Record Linkage: Relationships between File Sizes, Identifiers, and Match Weights. *Methods of Information in Medicine.* 2001;40:196-2003.
6. Marsolo, Keith. Approaches to Facilitate Institutional Review Board Approval of Multicenter Research Studies. *Medical Care.* 2012; 50(7):S77-S81.
7. Quantin C, Allaert F-A, Avillach P, et al. Building application-related patient identifiers: What solution for a European Country? *International Journal of Telemedicine and Applications.* 2008:1-5.
8. Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow up: quality assessment of an anonymous record linkage procedure. *International Journal of Medical Informatics.* 1998;49:117-122.
9. Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods of Information in Medicine.* 1998;37(3):271-277.
10. Schneier B. *Applied cryptography, protocols, algorithms, and source code.* Chichester: Wiley; 1994.
11. Dhir R, Patel AA, Winters S, Bisceglia M, Swanson D, Aamodt R, Becich MJ. A multidisciplinary approach to honest broker services for tissue banks and clinical data: a pragmatic and practical model. *Cancer.* 2008 Oct 1;113(7):1705-15.
12. Gladwell M. *The Tipping Point: How Little Things Can Make a Big Difference.* Boston, MA: Little, Brown & Company; 2000.

13. Stokols D, Hall KL, Taylor BK, Moser RP. The science of team science - Overview of the field and introduction to the supplement. *American Journal of Preventive Medicine*. 2008;35(2):S77-S89.
14. Hall K, Feng A, Moser R, Stokols D, Taylor B. Moving the science of team science forward: collaboration and creativity. *Am J Prev Med* 2008;35(2S):S243–S249: S161–S172.

## Chapter 4. An overview of record linkage methods

### Overview

Randomized controlled trials (RCTs) remain the gold standard for assessing intervention efficacy; however, RCTs are not always feasible or sufficiently timely. Perhaps more importantly, RCT results often cannot be generalized due to a lack of inclusion of “real world” combinations of interventions and heterogeneous patients (Clancy and Slutsky 2007; Office 2007; Smith 2007; Institute of Medicine 2009). With recent advances in information technology, data, and statistical methods, there is tremendous promise in leveraging ever-growing repositories of secondary data to support comparative effectiveness and public health research (Institute of Medicine 2009; Sox 2010; VanLare, Conway et al. 2010). While there are many advantages to using these secondary data sources, they are often limited in scope, which, in turn, limits their utility in addressing important questions in a comprehensive manner. These limitations can be overcome by linking data from multiple sources such as health registries and administrative claims data (Lipscomb, Gotay et al. 2005; Gliklich and Dreyer 2007; Bloomrosen M 2008; Brookhart, Sturmer et al. 2010).

Record linkage is an extensive and complex process that is both a science and an art (Roos, Wajda et al. 1986). While the process can be difficult to navigate, many effective strategies have been developed and documented in the health services literature. However, new policies and concerns over data security are making it more challenging for investigators to link data using traditional methods. Of particular importance, increasingly restrictive policies governing Protected Health Information (PHI) severely limits access to the unique identifiers on which many documented strategies rely. In light of these challenges, there is a need to increase understanding and extend capacity to perform reliable linkages in varying scenarios of data availability.

In this chapter, we guide the reader through an overview of data linkage methods and discuss the strengths and weaknesses of various linkage strategies in the effort to develop and document a set of best practices for conducting data linkages with optimal validity and reliability and minimal risk to privacy and confidentiality.

## Data cleaning and standardization

Data come in different shapes, sizes, and quality, creating scenarios that must be considered in building a linkage algorithm. For instance, demographic information often contains typographical and data entry errors, such as transposed SSN digits and misspellings. An individual's information sometimes changes over time, with life changes such as marriage or moving leading to changes in one's name or address. People sometimes deliberately report false information to defraud insurance providers or to avoid detection. Twins can have very similar information. Finally, the spouses and/or children of the family's primary health insurance subscriber sometimes use the primary subscriber's information. These idiosyncrasies are what make data linkage difficult, so the more work done upfront to clean and standardize the data, the better the chances of a successful linkage.

With this in mind, the first step after data delivery is to examine the nature of the data, paying particular attention to the way information is stored, the completeness of the identifying information, the extent to which information overlaps, and the presence of any idiosyncrasies in the data. By doing so, steps can be taken to clean and standardize the available information across data sources to minimize false matches attributable to typographical errors. Table 5 on the next page shows some of the commonly seen issues to clean, account for, and/or standardize before beginning the linkage process.

Many data manipulation techniques are available in commonly used software (e.g., SAS and Stata) to facilitate the data cleaning and standardization process. Using these techniques renders all linkage variables the same across data sources. That is, variables that will be compared are forced into the same case (e.g., all uppercase), the same format (e.g., 01SEP2013), the same content (e.g., stripped of all punctuation and digits), and the same length.

Additionally, identifiers can be parsed into separate pieces of information. For instance, full names can be parsed into first, middle, last, and maiden names; dates of birth can be parsed into month, day, and year of birth; and addresses can be parsed into street, city, state, and zip code. Parsing identifiers into separate pieces where possible allows the researcher to maximize the amount of available information and give credit for partial agreement when record pairs do not agree character-for-character. This is particularly important in accounting for changes across time, such as a name change after marriage or an address change after a move. In such cases, matching on the separate pieces allows for the possibility of

partial credit, which when combined with other information, may provide sufficient evidence that the records being compared represent the same person.

**Table 5. Common variations found in selected linkage identifiers**

Field	Type	Examples
Names	Case	John Smith   JOHN SMITH
	Nicknames	Charles   Chuck
	Synonyms	William   Bill
	Prefixes	Dr. John Smith
	Suffixes	John Smith, III
	Punctuation	O'Malley   Smith-Taylor   Smith, Jr.
	Spaces	John Smith, Jr
	Digits	John Smith
	Initials	AM   A.M.   Anne Marie
	Transposition	John Smith   Smith John
Address	Abbreviations	RD   Road   DR   Drive
Dates	Format	01012013   01-01-2013   01JAN2013
	Invalid values	Month = 13   Day = 32   Birth year = 2015
SSN	Format	999999999   999-99-9999   999 99 9999
Geographic Locations	Abbreviations	NC   North Carolina
	FIPS codes	North Carolina = 37
	SSA codes	North Carolina = 34
	Zip codes	99999   99999-9999
	Concatenation of state and county codes	Mecklenburg County, NC   37119
Sex	Format	Male / Female   M / F   1 / 2

Other techniques have been developed to account for minor misspellings and typographical errors. Here are three examples. 1) Strings can be converted to phonetic codes (e.g., SOUNDEX or NYSIIS) before comparison. 2) Strings can be compared using editing distance techniques to determine how many steps (i.e., insertions, deletions, transpositions, etc.) would be required to get from String A to String B (e.g., it would require one step – one character deletion – to get from “Billy” to “Bill”). 3) Names can be linked to an array of synonyms (e.g., “William” and “Bill”) to account for the use of nicknames.(Levenshtein 1966; Jaro 1989; Winkler 1990) Each of these techniques has been shown to improve the accuracy of string comparisons.(Jaro 1989; Winkler 1990) We have provided sample SAS code for performing these functions in Appendix 2.

Finally, it is important to consider informational overlap in the available linkage identifiers. As in statistical modeling where two correlated variables should not be included together in the model, variables with overlapping information should not be included in the same linkage algorithm. Otherwise, assigning credit for matches on both zip code and county is redundant, thereby overestimating the extent to which two records agree. Likewise, assigning separate penalties for non-matches on first name and first name initial is redundant, thereby overestimating the extent to which two records disagree. In the case of overlapping variables, researchers can pick either one or take the most informative match (e.g., if two records match on zip code and county, then the match on zip code is more informative).

How much data cleaning and standardization is necessary depends on the quality of the data, the research question, and the discretion of the researcher. While cleaning can improve linkage rates, the cleaning process can be quite labor-intensive, so researchers should consider the cost-benefit analysis before investing a significant amount of time on cleaning the data. If the data quality is poor and/or only a few identifiers are available, cleaning has been highly recommended (Wajda and Roos 1987). When the data quality is relatively good or many identifiers are available to override the errors, the cost usually outweighs the reward. (Randall SM 2013) The scope of the project should be considered as well. If the project is exploratory in nature, then the potential harm of false positive and false negatives may be negligible. Conversely, a hypothesis test may be significantly biased by false positives and false negatives (Krewski D 2005). Each of these considerations should be weighed against one another in the effort to determine whether data cleaning is prudent.

## **Linkage methods**

There are two main types of linkage algorithms: 1) deterministic and 2) probabilistic. Both have been successfully implemented in previous research studies (Newcombe and Kennedy 1962; Rogot, Feinleib et al. 1983; Rogot, Sorlie et al. 1986; Roos, Wajda et al. 1986; Clark 1993; Potosky, Riley et al. 1993; Bell, Keeseey et al. 1994; Clark and Hahn 1995; Jamieson, Roberts et al. 1995; Muse, Mikl et al. 1995; Doebbeling, Wyant et al. 1999; Cook, Olson et al. 2001; Blakely and Salmond 2002; Grannis, Overhage et al. 2003; Weiner, Stump et al. 2003; Winglee, Valliant et al. 2005; Li, Quan et al. 2006; Bradley, Given et al. 2007; Hammill, Hernandez et al. 2009; Jacobs, Edwards et al. 2010; Tromp, Ravelli et al. 2011; Nadpara 2012). Choosing the best algorithm to use in a given situation depends on many interacting factors, such as time; resources; the research question; and the quantity and quality of the variables available to link, including the degree to which they individually and collectively are able to

identify an individual uniquely. With this in mind, it is important that researchers are equipped with data linkage algorithms for varying scenarios. The key is to develop algorithms to extract and make use of enough meaningful information to make sound decisions. In this section, we will review the main algorithm types and discuss the strengths and weaknesses of each in an effort to derive a set of guidelines for which algorithms are best in varying scenarios of data availability, quality, and investigator goals.

### **Deterministic linkage methods**

Deterministic algorithms determine whether record pairs agree or disagree on a given set of identifiers, where agreement on a given identifier is assessed as a discrete – “all-or-nothing” – outcome. Match status can be assessed in a single step or in multiple steps. In a single step strategy, records are compared all at once on the full set of identifiers. A record pair is classified as a match if the two records agree, character-for-character, on all identifiers, and the record pair is uniquely identified (no other record pair matched on the same set of values). A record pair is classified as a non-match if the two records disagree on any of the identifiers, or if the record pair is not uniquely identified. In a multiple step strategy (also referred to as an iterative or stepwise strategy), records are matched in a series of progressively less restrictive steps in which record pairs that do not meet a first round of match criteria are passed to a second round of match criteria for further comparison. If a record pair meets the criteria in any step, it is classified as a match. Otherwise, it is classified as a non-match.

While the existence of a gold standard in registry-to-claims linkages is a matter of debate, the iterative deterministic approach employed by NCI to create the SEER-Medicare linked dataset (Potosky, Riley et al. 1993; Warren, Klabunde et al. 2002) has demonstrated high validity and reliability and has been employed successfully in multiple updates of the SEER-Medicare linked dataset (National Cancer Institute SEER-Medicare Program 2011; Warren and Carpenter August 20, 2010). The algorithm consists of a sequence of deterministic matches using different match criteria in each successive round.

In the first step, two records must match on SSN and one of the following:

- first and last name (allowing for fuzzy matches, such as nicknames);
- last name, month of birth, and sex; or
- first name, month of birth, and sex

If SSN is missing or does not match, or two records fail to meet the initial match criteria, they may be declared a match if they agree on the match criteria in a second round of deterministic

linkages, in which two records must match on last name, first name, month of birth, sex and one of the following:

- 7-8 digits of the SSN; or
- Two or more of year of birth, day of birth, middle initial, or date of death

In situations in which full identifiers or partial identifiers are available, but may not be released or transmitted, a deterministic linkage on encrypted identifiers may be employed. Quantin and colleagues (Fellegi and Sunter 1969; Quantin, Bouzelat et al. 1998; Quantin, Bouzelat et al. 1998) have developed procedures for encrypting identifiers using cryptographic hash functions so identifiers needed for linkage can be released directly to researchers without compromising patient confidentiality. A cryptographic hash function, such as the Secure Hash Algorithm version 2 (SHA-2) released by the National Security Agency (NSA) and published by NIST, is a deterministic procedure that takes input and returns an output that was intentionally changed by an algorithm. Due to its deterministic attributes and the inability to reverse engineer a hashed value, it is widely adapted in security applications and procedures. Quantin's research shows that records can successfully be linked via deterministic algorithms using identifiers encrypted before release (Quantin 1998 (1); Quantin, 1998 (2), Quantin 1998(3)). It should be noted, however, that record pairs matched on encrypted identifiers cannot be manually reviewed or validated.

### **Probabilistic linkage methods**

The deterministic approach ignores the fact that certain identifiers or certain values have more discriminatory power than others do. Probabilistic strategies have been developed to assess 1) the discriminatory power of each identifier, and 2) the likelihood that two records are a true match based on whether they agree or disagree on the various identifiers.

According to the model developed by Fellegi and Sunter (Fellegi and Sunter, 1969), matched record pairs can be designated as matches, possible matches, or non-matches, based on the calculation of linkage scores and the application of decision rules. Say we have two files, A and B, where file A contains 100 records and File B contains 1,000 records. The comparison space is the Cartesian product made up of all possible record pairs ( $A \times B$ ), or  $100 \times 1,000 = 100,000$  possible matches. Each pair in the comparison space is either a true match or a true non-match.

When dealing with large files (e.g., claims files), considering the entire Cartesian product is often computationally impractical. In these situations, it is advisable to reduce the comparison space to only

those matched pairs that meet certain basic criteria. This is referred to as “blocking,” which serves to functionally subset a large dataset into a smaller dataset of individuals with at least one common characteristic, such as their geographic region, or a specific clinical condition. For instance, the number of matched pairs to be considered may be limited to only those matched pairs that agree on clinical diagnosis, or on month of birth and county of residence. Those record pairs that do not meet the matching criteria specified in the blocking phase are automatically classified as non-matches and removed from consideration. In many instances, one pass will not capture all possible matches, so multiple passes or blocks are necessary to avoid automatic misclassification. Since two records cannot be matched on missing information, the variables chosen for the blocking phase should be relatively complete, having few missing values. Blocking strategies such as this reduce the set of potential matches to a more manageable number. Because blocking strategies can influence linkage success, Christen and Goiser recommend that researchers report the specific steps of their blocking strategy (Christen and Goiser, 2007).

For every matched pair identified in the blocking phase, the two records are compared on each linkage identifier, producing an agreement pattern. The weight assigned to agreement or disagreement on each identifier is assessed as a likelihood ratio, comparing the probability that true matches agree on the identifier (“*m*-probability”) to the probability that false matches randomly agree on the identifier (“*u*-probability”).

The *m*-probability can be estimated based on values reported in published linkage literature, or by taking a random sample of pairs from the comparison space, assigning match status via manual review, and calculating the probability that two records agree on a particular identifier when they are true matches. The *u*-probability can be calculated by observing the probability that two records agree on a particular identifier merely by chance (e.g., the *u*-probability for month of birth is 1/12, or .083). Calculating value-specific *u*-probabilities for an identifier based on the frequency of each value and the likelihood that two records would agree on a given value simply by chance yields additional information. For instance, a match on a rare surname such as “Lebowski” is less likely to occur by chance, and is thereby assigned greater weight than a match on a common surname such as “Smith”. This lesson can be applied to any linkage identifier for which values are differentially distributed.

When two records agree on an identifier, an agreement weight is calculated by dividing the  $m$ -probability by the  $u$ -probability, and taking the  $\log_2$  of the quotient. For example, if the probability that true matches agree on month of birth is 97% and the probability that false matches randomly agree on month of birth is 8.3% (1/12), then the agreement weight for month of birth would be calculated as  $\log_2(.97/.083)$ , or 3.54. When two records disagree on an identifier, a disagreement weight is calculated by dividing 1 minus the  $m$ -probability by 1 minus the  $u$ -probability. For example, the disagreement weight for month of birth would be calculated as  $\log_2(1-.97) / \log_2(1-.083)$ , or - 4.93.

While the method above accounts for the discriminatory power of the identifier, it does not yet take into account the *degree* to which records agree on a given identifier. Assigning partial agreement weights in situations where two strings do not match character-for-character can account for minor typographical errors, including spelling errors in names or transposed digits in dates or SSNs. Partial agreement weights can be assigned by converting an identifier to a string and using string comparator methods (Levenshtein 1966, Jaro 1989, Winkler 1990) to determine the probability that two strings match. The full agreement weight for the identifier can then be multiplied by the calculated probability that the two strings match to generate a partial agreement weight that is proportional to the confidence that the two strings are a match. For example, if the full agreement weight for first name is 12, and the string comparator method indicates that there's a 95% probability that the first name on one record matches the first name on another record, then the partial agreement weight would be equal to  $12 * 0.95$ , or 11.4). Once the weights, full and partial, for each identifier have been calculated, the linkage score for each matched pair is equal to the sum of the weights across all linkage identifiers.

An initial assessment of linkage quality can be gained by plotting the match scores in a histogram. If the linkage algorithm is working properly, then the plot should show a bimodal distribution of scores, with one large peak among the lower scores for the large proportion of likely non-matches and a second smaller peak among the higher scores for the smaller set of likely matches. The cutoff threshold for match/non-match status will be a score somewhere in the trough between the two peaks. Depending on the research question and the nature of the study, the initial threshold can be adjusted to be more conservative (higher score) or more liberal (lower score). A more conservative threshold will maximize the accuracy of the linkage decision, as only those record pairs with a high score will be counted as matches. Conversely, a more liberal threshold will maximize the sensitivity of the linkage decision to possible matches.

Cook and colleagues (Cook, Olson et al. 2001) define the cutoff threshold as the difference between the desired weight and the starting weight. Given two files, A and B, the starting weight for each record pair is equal to the log<sub>2</sub> of the odds of picking a true match by chance,

$$\log_2( E / ((A \times B) - E))$$

where E is the number of expected matches, A is the number of records in File A and B is the number of records in File B. If the number of records in File A is 1,000, the number of records in File B is 1,000,000, and the expected number of matches is 10%, or 100, then the starting weight will be -19.93. The “expected” number of matches can be determined by prior research, prior knowledge or an educated guess if there is no precedent.

The desired weight for each record pair is equal to the log<sub>2</sub> of the odds associated with the desired probability that two records were not matched together by chance is equal to,

$$\log_2( P / (1 - P))$$

where P is the desired probability that two records were not matched merely by chance. If the desired probability is 0.95, then the desired weight is  $\log_2( 0.95 / (1 - 0.95)) = 4.25$ . The cutoff threshold, or score needed to have a false match rate of <0.05, is the difference between the desired weight, 4.25, and the starting weight, -19.93, which is 24.18. If the computed linkage score is greater than or equal to the cutoff threshold, then the record pair is classified as a match. If the computed linkage score is less than the cutoff threshold, then the record pair is classified as a non-match. Researchers wishing to maximize the sensitivity of the algorithm to potential matches can relax this threshold somewhat and manually review all record pairs with scores near the calculated cutoff.

Once this process is complete, a sample of the match decisions made by the linkage algorithm should be reviewed to ensure that the algorithm performed as intended. By reviewing match decisions, you can often identify conditions in which the algorithm could use some tweaking to account for difficult cases. For instance, frequently the children and/or spouses of the primary subscriber use the primary subscriber’s SSN, thereby making it difficult to identify them as unique individuals given the large weight often assigned to agreement on SSN. Twins are another difficult case, as they have the same birthdate, frequently have similar names, and often have SSNs that differ on only 1-2 digits. Records for a married woman can sometimes be difficult to match when marriage leads to changes in the woman’s

last name and/or address. Finally, tweaks to the algorithm can often improve the performance of the algorithm on the “fuzzy” or borderline cases. By reviewing a sample of match decisions, you can tweak your algorithm to account for each of these cases, thereby improving the performance of the algorithm. This will be particularly important for researchers who hope to re-use the algorithm for future linkages (e.g., matching a new year of registry cases to a new year of administrative claims data).

## Alternative methods

The methods presented above are those most commonly used in registry-to-claims linkages. Other methods are available for researchers who have more challenging linkage scenarios. The EM algorithm (Jaro 1989, Winkler) is an iterative approach to estimating  $m$ - and  $u$ -probabilities. According to Winkler<sup>49</sup>, the EM algorithm provides very accurate estimates of  $m$ - and  $u$ -probabilities in situations where the amount of typographical error in the identifiers is minimal, but performs poorly when the identifiers contain numerous typographical errors. In the sorted-neighborhood approach (Hernandez and Stolfo, 1995), the data sources are stacked and sorted by various combinations of the available identifiers. For each sort, all records within a window of  $n$ -records are compared. The Bayesian approach (Belin and Rubin, 1995) is an alternative to the frequentist approach presented above. The computer science literature also includes distance-based techniques (Dey, Sarkar and De, 1998) as well as unsupervised and supervised machine learning approaches (Cochinwala, Kurien, et.al, 1998; Verykios and Moustakides, 2001).

## Selecting a linkage method

What kind of linkage method to employ in a given situation depends on a variety of factors, some of which are scientific and some of which are more subjective. In information-rich scenarios where direct identifiers are available and of good quality, deterministic methods have been recommended (Roos and Wajda, 1991). In these scenarios, deterministic methods are easy to implement, easy to interpret, and effective. In non-information-rich and information-poor scenarios in which direct identifiers are unavailable and/or the data are of a poor quality, probabilistic methods consistently outperform deterministic methods and thus validate the extra time and resources required to implement them. Beyond these broad guidelines, the decision is left to the researcher and his/her goals for the project. Herein lies the “art” of record linkage.(Roos, Wajda et al. 1986) For instance, researchers studying a rare disease may want to employ probabilistic methods even in information-rich scenarios in the effort to identify every possible match and maximize sample size. Ultimately, every researcher must weigh the

pros and cons of the available methods in the context of the project, and choose the method that best fits the budget, timeline, allotted resources, and research question.

## Evaluating linkage algorithms

In record linkage, there are two types of accuracy errors. A Type I error occurs when a true non-match is classified as a match. A Type II error occurs when a true match is classified as a non-match. Minimizing these errors is critical, particularly when the product will be used in a cohort study (Krewski, Wang, et.al, 2005), where linkage error can introduce bias into analyses. In the health sciences literature, the four metrics most often used to evaluate the accuracy of a linkage algorithm are: 1) sensitivity, 2) specificity, 3) positive predictive value, and 4) negative predictive value. Table 6 shows all possible outcomes of a linkage decision.

**Table 6. True match status by algorithm output**

		<b>ALGORITHM OUTPUT</b>	
		<b>MATCH</b>	<b>NON-MATCH</b>
<b>TRUE MATCH STATUS</b>	<b>MATCH</b>	A	B
	<b>NON-MATCH</b>	C	D

“A” represents all true matches that are correctly classified as matches; “B” represents all true matches that are incorrectly classified as non-matches; “C” represents all true non-matches that are incorrectly classified as matches; and “D” represents all true non-matches that are correctly classified as non-matches.

Sensitivity =  $A / A + B$ . Sensitivity measures the ability of an algorithm to correctly classify true matches as matches.

Specificity =  $C / C + D$ . Specificity measures the ability of an algorithm to correctly classify true non-matches as non-matches.

Positive predictive value (PPV) =  $A / A + C$ . PPV represents the proportion of matched pairs classified by the algorithm as matches that are true matches.

Negative predictive value (NPV) =  $D / B + D$ . NPV represents the proportion of matched pairs classified by the algorithm as non-matches that are true non-matches.

These metrics measure an algorithm’s ability to classify correctly true matches as matches, and true non-matches as non-matches.

Due to the large number of potential matches identified during the blocking phase, the bulk of the comparison space will be made up of true non-matches. For this reason, Christen and Goiser (Christen and Goiser, 2007) argue that linkage metrics that include true non-matches (e.g., specificity and negative predictive value) in the equation will be skewed. Instead, they recommend metrics such as the *f-measure* (van Rijsbergen, 1979) that represents the harmonic mean of the sensitivity, and positive predictive value that is not influenced by the large number of true non-matches. The f-measure is calculated as

$$((\beta^2 + 1.0) * \text{Sensitivity} * \text{PPV}) / (\beta^2 * \text{Sensitivity} + \text{PPV})$$

where  $\beta$  is equal to the user-assigned relative importance of sensitivity over positive predictive value. If the user wishes to assign equal weight to sensitivity and positive predictive value, then  $\beta = 1.0$ . If the user wishes to assign sensitivity twice the weight of positive predictive value, then  $\beta = 2.0$ .

Trade-offs between metrics should be expected. Emphasis on PPV often leads to sacrifices in sensitivity, while emphasis on sensitivity often leads to sacrifices in PPV. While there is no hard rule, a good linkage algorithm will typically have values in excess of 95% across all metrics. However, what is acceptable depends on the context of the study. If the study involves the testing of hypotheses and/or the results will have significant practical implications (e.g., findings will be incorporated into clinical practice or influence policy), a higher percent match is more desirable. On the other hand, if a study is exploratory, a lower percent match may be acceptable. What is acceptable may also vary depending on the frequency of the outcome. Researchers studying a rare disease may seek to emphasize sensitivity to maximize the sample size, while a researcher studying a more frequently occurring disease may want to maximize PPV to ensure that matches identified by the algorithm are true matches.

## **Validating linkage results**

The final step of the linkage process is the validation of the match results. Initial steps for determining linkage validity are to look for ties in which multiple record pairs identified as matches by the algorithm have the exact same set of values (e.g., common names such as “Mary Smith” or “Mike Brown” are typically repeated in large datasets, and thus have multiple matches). Where possible, ties should be adjudicated by reference to additional information. If no additional information is available, then the record pairs should be classified as non-matches. If an algorithm is successful, there will be few to no ties.

The next step is to assess the extent to which your matched sample reflects the target population. For instance, in a study linking a single state's cancer registry to Medicare administrative claims for that state, researchers may use estimates of the percentage of cancer patients over the age of 65 to determine what percentage of patients in the cancer registry would be expected to be linked to the Medicare data. If estimates indicate that 60% of cancer patients in the state are over 65, then it is reasonable to expect that 60% of the patients in the cancer registry will be matched with Medicare. If, instead, the researcher finds that only 30% of patients in the cancer registry are successfully matched, this may serve as a signal that there is a problem with the matching algorithm.

While not well documented in the literature, some form of manual review is typically employed to check the results. Before starting the manual review process, a set of decision rules is developed to standardize the decision process across reviewers. Next, a random sample is drawn from the set of all potential matches identified during the blocking phase. Following the decision rules, one or more reviewers then determine whether each potential match is a match or non-match. Finally, the decisions documented during the manual review process are used as a gold standard against which the decisions made by the algorithm are compared, allowing for the calculation of the sensitivity, positive predictive value, and f-measure of the algorithm. A good algorithm should have scores of 95% or better across the four metrics.

## **Final remarks**

In this chapter, we have provided an overview of data linkage methodology, from the point of data delivery to the reporting of the linkage results. We began by examining the quality of the data. Here we found that if the data quality is good (e.g., unique identifiers with few typographical errors and/missing values), the cost of the labor-intensive process of cleaning and standardizing the data is not worth the reward, and therefore not recommended. In this scenario, deterministic linkage methods are accurate, straightforward, and easy to implement.

If the data quality is poor (e.g., little identifying information and/or numerous typographical errors), cleaning and standardizing the data before linkage can greatly improve linkage rates. In these scenarios, iterative deterministic techniques or more-sophisticated probabilistic techniques are recommended. Combining the two methods can improve efficiency and save computational resources. Using this method, a deterministic match on all identifiers can be executed first to identify certain matches. The remaining record pairs that disagree on at least one of the identifiers can be submitted for probabilistic

matching. This method reduces the number of record pairs that will be processed in the more resource-intensive probabilistic matching phase.

In order to limit the computational resources required to compare the Cartesian product of all possible matches, blocking should be implemented to reduce the comparison space to record pairs that agree on some basic criteria (e.g., date of birth and county, clinical diagnosis, etc.). This process can improve computational efficiency and performance substantially. Blocking techniques can have a significant effect on the linkage results, and thus, researchers should report the blocking method.

Both the deterministic and probabilistic procedures should be considered iterative. After completing the initial linkage, a random sample of match decisions should be reviewed to ensure that the algorithm is performing as intended. If the review process reveals opportunities for improvement, then the algorithm should be adjusted to account for the identified weaknesses.

Once the linkage process is complete, the results should be compared to known metrics. For instance, if it is known that 20% of cancer patients in the state are covered by private insurance, then roughly 20% of the records in a state cancer-registry database should match to private insurance claims. If the observed match rate differs substantially from the expected results, then the linkage method should be re-evaluated and repeated.

When reporting linkage results, estimates of the sensitivity, positive predictive value, and f-measure of the algorithm should be reported to provide readers with a characterization of the validity and reliability of the linkage product. Due to the disproportionately large number of non-matched pairs identified during the blocking phase, measures that include the number of non-matched pairs in the calculation (e.g., specificity and negative predictive value) should not be reported.

In the next chapter, we will empirically demonstrate the approaches described in this chapter, and develop and test a series of deterministic and probabilistic algorithms for scenarios of varying unique identifier availability.

Researchers who wish to learn more about data linkage approaches and techniques beyond those covered in this chapter are referred to Dr. William Winkler's list of Statistical Data Editing

References:<http://citeseerx.ist.psu.edu/viewdoc/summary;jsessionid=BAA7B495D9CFBEB3276C67AB96BFFA6D?doi=10.1.1.79.1519>

## References

1. Clancy CM, Slutsky JR. Commentary: a progress report on AHRQ's Effective Health Care Program. *Health services research*. Oct 2007;42(5):xi-xix.
2. Office CB. Research on the Comparative Effectiveness of Medical Treatments: Issues and Options for an Expanded Federal Role. *Pub. No. 2975 Washington DC*. 2007.
3. Smith S. Preface. *Medical Care*. 2007;45(10 Suppl 2):S1-S2.
4. Institute of Medicine. Initial National Priorities for Comparative Effectiveness Research. *Washington DC: National Academics Press*. 2009.
5. Sox HC. Comparative effectiveness research: a progress report. *Ann Intern Med*. Oct 5 2010;153(7):469-472.
6. VanLare JM, Conway PH, Sox HC. Five next steps for a new national program for comparative-effectiveness research. *N Engl J Med*. Mar 18 2010;362(11):970-973.
7. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Medical Care*. Jun 2010;48(6 Suppl):S114-120.
8. Lipscomb J, Gotay C, Snyder C. *Outcomes Assessment in Cancer: Measures, Methods, and Applications*. Cambridge: Cambridge University Press; 2005.
9. Gliklich R, Dreyer NA. *Registries for Evaluating Patient Outcomes: A User's Guide*. (AHRQ Publication No. 07-EHC001-1). Rockville, MD: Agency for Healthcare Research and Quality; 2007.
10. Bloomrosen M DD. Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association. *J Am Med Inform Assoc*. 2008;15(6):715-722.
11. Roos LL, Wajda A, Nicol JP. The art and science of record linkage methods that work with few identifiers. *Comput Biol Med*. 1986;16(1):45-57.
12. Levenshtein V. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*. 1966;10:707-710.
13. Jaro MA. Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*. 1989;84(406):414-420.

14. Winkler WE. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *American Statistical Association, Proceedings of the Section on Survey Research Methods*. 1990.
15. Wajda A, Roos LL. Simplifying record linkage: software and strategy. *Comput Biol Med*. 1987;17(4):239-248.
16. Randall SM FA, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making*. 2013;13.
17. Krewski D DA, Wang Y, Bartlett S, Zielinski JM, Mallick R. The effect of record linkage errors on risk estimates in cohort mortality studies. *Survey Methodology*. 2005;31(1):13-21.
18. Doebbeling BN, Wyant DK, McCoy KD, et al. Linked insurance-tumor registry database for health services research. *Medical Care*. Nov 1999;37(11):1105-1115.
19. Jacobs JP, Edwards FH, Shahian DM, et al. Successful linking of the society of thoracic surgeons adult cardiac surgery database to centers for medicare and medicaid services medicare data. *Annals of Thoracic Surgery*. 2010;90:1150-1157.
20. Li B, Quan H, Fong A, Lu M. Assessing record linkage between health care and vital statistics databases using deterministic methods. *BMC Health Serv Res*. 2006;6(1):1-10.
21. Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. *Stat Med*. Mar 15-Apr 15 1995;14(5-7):499-509.
22. Nadpara PA. Linking Medicare, Medicaid, and Cancer Registry Data to Study the Burden of Cancers in West Virginia. *Medicare & Medicaid Research Review*. 2012;2(4).
23. Weiner M, Stump TE, Callahan CM, Lewis JN, McDonald CJ. A practical method of linking data from medicare claims and a comprehensive electronic medical records system. *International Journal of Medical Informatics*. 2003;71:57-69.
24. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *International journal of epidemiology*. 2002;31:7.
25. Cook LJ, Olson LM, Dean JM. Probabilistic Record Linkage: Relationships between File Sizes, Identifiers, and Match Weights. *Methods of Information in Medicine*. 2001;40:196-2003.
26. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a Probabilistic Record Linkage Technique without Human Review. *AMIA 2003 Symposium Proceedings*. 2003:259-263.
27. Newcombe HB, Kennedy JM. Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM*. 1962;5(11):563-566.

28. Rogot E, Feinleib M, Ockay KA, Schwartz SH, Bilgrad R, Patterson JE. On the feasibility of linking census samples to the National Death Index for epidemiologic studies: a progress report. *Am J Public Health*. Nov 1983;73(11):1265-1269.
29. Rogot E, Sorlie P, Johnson NJ. Probabilistic methods in matching census samples to the National Death Index. *Journal of Chronic Disease*. 1986;39(9):719-734.
30. Bell RM, Keesey J, Richards T. The urge to merge: linking vital statistics records and Medicaid claims. *Medical care*. Oct 1994;32(10):1004-1018.
31. Bradley CJ, Given CW, Luo Z, Roberts C, Copeland G, Virnig BA. Medicaid, Medicare, and the Michigan Tumor Registry: a linkage strategy. *Med Decis Making*. Jul-Aug 2007;27(4):352-363.
32. Clark DE. Development of a statewide trauma registry using multiple linked sources of data. *Proc Annu Symp Comput Appl Med Care*. 1993:654-658.
33. Clark DE, Hahn DR. Comparison of Probabilistic and Deterministic Record Linkage in the Development of a Statewide Trauma Registry. *American Medical Informatics Association*. 1995:5.
34. Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. *Am Heart J*. 2009;157(6):995-1000.
35. Jamieson E, Roberts J, Browne G. The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. *Methods of Information in Medicine*. 1995;34:371-377.
36. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. *Journal of clinical epidemiology*. May 2011;64(5):565-572.
37. Winglee M, Valliant R, Scheuren F. A case study in record linkage. *Survey Methodology*. 2005;31(1):3-11.
38. Potosky AL, Riley GF, Lubitz JD, Mentnech RM, Kessler LG. Potential for cancer related health services research using a linked medicare-tumor registry database. *Medical Care*. 1993;31(8):732-748.
39. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical Care*. Aug 2002;40(8 Suppl):IV-3-18.

40. National Cancer Institute SEER-Medicare Program. Search SEER-Medicare Publications. 2011; <http://healthservices.cancer.gov/seermedicare/overview/publications.html>. Accessed March 4, 2011.
41. Warren J, Carpenter W. Email: Details on SEER-Medicare linkage methods. August 20, 2010.
42. Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow up: quality assessment of an anonymous record linkage procedure. *International Journal of Medical Informatics*. 1998;49:117-122.
43. Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods of Information in Medicine*. 1998;37(3):271-277.
44. Quantin C, Allaert F-A, Avillach P, et al. Building application-related patient identifiers: What solution for a European Country? *International Journal of Telemedicine and Applications*. 2008:1-5.
45. Fellegi IP, Sunter AB. A Theory for Record Linkage. *Journal of the American Statistical Association*. 1969;64(328):1183-1210.
46. Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication. *Studies in Computational Intelligence*. 2007; 43:127-151.
47. Roos LL, Wajda A. Record linkage strategies. Part 1: estimating information and evaluating approaches. *Methods of Information in Medicine*. 1991; 30:117-123.
48. van Rijsbergen CJ. *Information retrieval: data structures and algorithms*. London: Butterworths, 1979.
49. Winkler WE. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 667-671.
50. Hernandez MA, Stolfo SJ. The merge/purge for large databases, Proceedings of the SIGMOD 95 conference, 1995, pp. 127-138.
51. Belin TR, Rubin DB. A method for calibrating false matches in record linkage. *Journal of the American Statistical Association*. 1995; 90: 694-707.
52. Dey D, Sarkar S, De P. Entity matching in heterogenous databases: A distance based decision model, Proceedings of the 31<sup>st</sup> Hawaii International Conference on System Sciences, 1998.
53. Cochinwala M, Kurien V, Lalk G, Shasha D. Efficient data reconciliation. Bellcore, 1998.

54. Verykios VS, Moustakides GV. A cost optimal decision model for record matching, "Workshop on Data Quality: Challenges for Computer Science and Statistics, 2001.
55. Campbell KM, Deck D, Krupski AK. Record linkage software in the public domain: A comparison of Link Plus, the Link King, and a basic deterministic algorithm. *Health Informatics Journal*. 2008; 14: 5-15.
56. Day C. Record linkage I: Evaluation of commercially available record linkage software for use in NASS. *National Agriculture Statistics Service Research Division technical report STB Research Report Number STB-95-02*, 1995.
57. William WE. Record Linkage Software and Methods for Merging Administrative Lists. *Statistical Research Division Technical Report RR01|03*, U.S. Bureau of Census, 2001.
58. Jones L, Sujansky W. Patient data matching software: A buyers guide for the budget conscious. California Health Care Foundation, 2004.

## Appendix 2. Useful SAS functions and procedures

### Data cleaning and standardization

#### *Remove all special characters*

**Syntax:** `compress(varname, "., ' {} [] / \ ( ) + ~ ` ! @ # $ % ^ & * _ < > ?");`

**Ex.** Remove all special characters from the string "(John\_Doe!)"  
`newname = compress(name, "., ' {} [] / \ ( ) + ~ ` ! @ # $ % ^ & * _ < > ?");`  
`put name;`  
 JohnDoe

#### *Remove a specific character*

**Syntax:** `compress(varname, "-");`

**Ex.** Remove all dashes from the string "999-99-9999"  
`newssn = compress(ssn, "-");`  
`put newssn;`  
 999999999

#### *Remove all punctuation*

**Syntax:** `compress(varname, "P");`

**Ex.** Remove all punctuation from "O'Brien-Smith"  
`newname = compress(name, , "P");`  
`put newname;`  
 OBrienSmith

#### *Remove all digits*

**Syntax:** `compress(varname, "D");`

**Ex.** Remove all digits from the string "J2ohn"  
`newname = compress(name, , "D");`

```
put newname;
John
```

### ***Remove all spaces***

**Syntax:** `compress(varname, " ");`

**Ex.** Remove all spaces from the string "Van Slyke"

```
newname = compress(name, " ");
put newname;
VanSlyke
```

### ***Remove leading and trailing spaces***

**Syntax:** `strip(varname);`

**Ex.** Remove all leading and trailing spaces from the string " John "

```
newname = strip(firstname);
put newname;
John
```

### ***Remove extra spaces with a single space***

**Syntax:** `combl(varname);`

**Ex.** Remove all extra spaces from the string "Van Slyke"

```
newname = compbl(name);
put newname;
Van Slyke
```

### ***Substring***

**Syntax:** `substr(varname, start, count);`

**Ex.** Extract first initial from the name "John"

```
initial = substr(name, 1, 1);
put initial;
J
```

***Search for and replace a string***

**Syntax:** tranwrđ(varname, string\_to\_be\_replaced, replacement\_string);

**Ex.** Find all instances of the string "Road" in the string "Anystreet Road" and replace with "Rd."

```
newstreet = tranwrđ(street, "Road", "Rd.");
put newstreet;
Anystreet Rd.
```

***Make all characters in a string uppercase***

**Syntax:** upcase(varname);

**Ex.** Make all characters in the string "John" uppercase

```
newname = upcase(name);
put newname;
JOHN
```

***Make all characters in a string lowercase***

**Syntax:** lowercase(varname);

**Ex.** Make all characters in the string "JOHN" lowercase

```
newname = lowercase(name);
put newname;
john
```

***Make all characters in a string proper case (1<sup>st</sup> character uppercase, remaining characters lowercase)***

**Syntax:** propcase(varname);

**Ex.** Convert the string "JOHN" to proper case

```
newname = propcase(name);
put newname;
```

John

### *Concatenate two strings*

**Syntax:** varname1||varname2;

-OR-

**Syntax:** cat(varname1,varname2);

**Ex.** Concatenate first name ("JOHN") and last name ("DOE")

```
newname = firstname||lastname;
newname = cat(firstname, lastname);
put newname;
JOHNDOE
```

\*\*\* For more information on CAT functions see:  
 "Purrfectly fabulous feline functions" by Louise S. Hadden in NESUG

### *Parse out pieces in a string*

**Syntax:** scan(varname, count, delimiter);

**Ex.** Parse out first, middle and last name from "JOHN SMITH DOE"

```
firstname = scan(fullname, 1, " ");
put firstname;
JOHN

middlename = scan(fullname, 2, " ");
put middlename;
SMITH

lastname = scan(fullname, 3, " ");
put lastname;
DOE
```

### *Convert string to phonetic code*

**Syntax:** soundex(varname);

**Ex.** Convert the first name "JOHN" to phonetic code

```

newname = soundex(firstname);
put newname;
J5

```

### *Calculate editing distance*

**Syntax:** `spedis(varname);`

-OR-

**Syntax:** `complev(varname);`

-OR-

**Syntax:** `compged(varname);`

**Ex.** Determine the editing distance between 'CHARLES' and 'CHARLIE'

```
name1 = 'CHARLES';
```

```
name2 = 'CHARLIE';
```

```
spedis = spedis(name1, name2);
```

```
put spedis;
```

```
21
```

```
complev = complev(name1, name2);
```

```
put complev;
```

```
2
```

```
compged = compged(name1, name2);
```

```
put compged;
```

```
200
```

\*\*\* Note: For a comparison of the editing distance functions, see "Fuzzy Matching using the COMPGED Function" by Paulette Staum in Northeast SAS Users Group (2007)

### **SSN validation**

See "Identifying Invalid Social Security Numbers" by Paulette Staum and Sally Dai in the Northeast SAS Users Group (2007)

### **Variable encryption**

**Syntax:** `md5(varname);`

**Ex.** Encrypt the name "JOHNDOE"

```
newname = md5(name);
```

```
put newname;
```

```
''ãiEfúÀ¥-q...zÅ
```

## General matching techniques

### *Loop to determine how many SSN digits match*

```
attrib SSN_DIG_MATCH format = 1. label = '# of SSN Digits that Match';
SSN_DIG_MATCH = 0;

do i = 1 to 9;
  if substr(put(CCR_SSN,$9.),i,1) = substr(put(BCBS_SSN,$9.),i,1) then
    SSN_DIG_MATCH + 1;
end;
```

### *Compare up to length of shortest string*

**Syntax:** compare(compress(string1,' '), compress(string2,' '), ':');

**Ex.** Compare the last names 'WILLIAMS-SMITH' and 'WILLIAMS' character-for-character up to length of shortest name

```
if compare(compress(lastname1,' '), compress(lastname2,' '), ':') = 0
then ln_compare = 'match';
```

\*\*\* Use with caution. This technique is a useful way to identify cases to manually review (e.g., names reported differently over time).

### *Compare strings in reverse order*

**Syntax:** compare(compress(reverse(string1),' '), compress(reverse(string2),' '), ':');

**Ex.** Compare the last names 'WILLIAMS-SMITH' and 'SMITH' character-for-character up to length of shortest name

```
if compare(compress(reverse(lastname1),' '),
compress(reverse(lastname2),' '), ':') = 0
then ln_compare = 'match';
```

\*\*\* Use with caution. This technique is a useful way to identify cases to manually review (e.g., names reported differently over time).

## SEER-Medicare algorithm

```
/* INITIATE SEER-MEDICARE MATCH INDICATOR */
attrib seermedicare format = 1. length = 3.;
seermedicare = 0;
```

```

/* IF SSN MATCHES */
if ssn then do;

    /* IF FIRST NAME AND LAST NAME MATCH */
    if ( firstname and lastname )

    or

    /* IF LAST NAME, MONTH OF BIRTH AND GENDER MATCH */
    (lastname and birthmonth and gender )

    or

    /* IF FIRST NAME, MONTH OF BIRTH AND GENDER MATCH */
    ( firstname and birthmonth and gender )
    then seermedicare = 1;
end;

/* IF SSN DOES NOT MATCH, AND LAST NAME, FIRST NAME, MONTH OF BIRTH */
/* AND GENDER MATCH */
if not ssn and ( lastname and firstname and birthmonth and gender )
then do;

    /* IF 7+ SSN DIGITS MATCH OR 2+ OF YEAR OF BIRTH, DAY OF BIRTH, */
    /* MIDDLE INITIAL OR DATE OF DEATH MATCH */
    if (ssn_dig_match >= 7) or
        (sum(of birthyear, birthday, middleinitial, deathdate) >= 2)
    then seermedicare = 1;
end;

```

## Probabilistic matching techniques

### *Calculate u-probability*

**Ex.** Generate *u*-probability for firstname

```

proc freq data = lib.dataset_a (keep = firstname) noprint;
table firstname / norow nocol out = lib.fnfreq;
run;

data lib.dataset_a (keep = firstname fn_uprob);
set lib.dataset_a;
attrib fnu length = 8. format = 10.9;
fn_uprob = percent / 100;
run;

```

### Calculate probabilistic weights

```

/* MATCH FIRST NAMES */
if firstname1 = firstname2 then fn_match = 'match';
else fn_match = 'nonmatch';

/* CALCULATE AGREEMENT WEIGHT FOR FIRST NAME AS: */
/* LOG2 (M_PROB/U_PROB) */
/* ASSUME M-PROBABILITY FOR FIRST NAME = 0.95 */
if fn_match = 'match' then do;
  fn_agree_weight = log2(0.95 / fn_uprob);
end;

/* CALCULATE DISAGREEMENT WEIGHT FOR FIRST NAME AS: */
/* LOG2( (1-M_PROB) / (1-U_PROB) ) */
/* ASSUME M-PROBABILITY = 0.95 */
else if fn_match = 'nonmatch' then do;
  fn_disagree_weight = log2( (1-0.95) / (1-fn_uprob) );
end;

```

### Use editing distance method to calculate probability that 2 strings are a match

```

/* IF PROB. THAT 2 STRINGS ARE A MATCH > .95, THEN IT'S A MATCH */
if mean(
  1 - (length(firstname1) * spedis(firstname1, firstname2)) / 2400),
  1 - (length(firstname2) * spedis(firstname2, firstname1)) / 2400)
  >= .95
then firstname_spedis = 'match';

*** Source: "Fuzzy Key Linkage" by Sigurd Hermansen in Southeast SAS Users
Group (2001)

```

### Efficiency and optimization techniques

#### Syntax:

```

proc sql;
create index varname on datasetname (varname);
quit;

```

**Ex.** Create index on SSN to improve performance of match on SSN

```

proc sql;
create index ssn on dataset_a (ssn);

```

```
quit;  
  
options msglevel = I;  
proc sql;  
create table matches as  
select *  
from dataset_a as f1, dataset_b as f2  
where f1.ssn = f2.ssn;  
quit;
```

### Appendix 3. Data linkage software packages

Manually writing the code to perform each step of the data linkage process in software packages such as SAS, MS SQL Server, or R gives the user full control over the entire process. While manual coding is ideal, the person writing the code must be familiar with linkage theory and must possess a great deal of programming expertise, qualities that may require funding for an expensive programmer. Furthermore, the amount of time required to write the code and the amount of computer resources needed to execute the linkage can be substantial. For researchers who lack the time, computer resources, expertise or personnel required to write the needed code manually, many publicly and commercially available products to streamline and simplify the linkage process are available. Here, we present several commonly used products.

#### Publically available packages

The following data linkage packages are available at no charge to the public:

**Link Plus**, developed by the Centers for Disease Control, is available at [www.cdc.gov/cancer/registryplus/lp.html](http://www.cdc.gov/cancer/registryplus/lp.html). The package provides a graphical user interface that is

straightforward and easy to use, requiring only beginner-level knowledge of the linkage process. In an evaluation performed by Campbell et al, Link Plus had an aggregate PPV of 94.6 and an aggregate sensitivity of 94.1 (for probabilistic scores between 16 and 25) (Campbell, et. al., 2008). While Link Plus is easy to use, it currently struggles with datasets with > 1 million records, which makes it difficult for researchers attempting to link claims datasets that typically contain millions of records.

**The Link King**, developed by Washington State's Division of Alcohol and Substance Abuse, is available at [www.the-link-king.com](http://www.the-link-king.com). The product itself is free, but it requires a license for base SAS, which currently costs ~ \$2,000. Like Link Plus, the package provides a graphical user interface that is straightforward and easy to use, requiring only beginner-level knowledge of the linkage process. In the evaluation performed by Campbell, the Link King had an aggregate PPV of 96.1 and an aggregate sensitivity of 96.7 (for certainty level 4) (Campbell, et. al., 2008). While the Link King is capable of handling larger datasets, it requires first and last name, as well as SSN or date of birth, which means that it can be used only in information-rich scenarios.

**ChoiceMaker 2** (developed by ChoiceMaker Technologies and available at [www.sourceforge.net/projects/oscm/](http://www.sourceforge.net/projects/oscm/)) and **FEBRL** (developed by the ANU Data Mining Group and available at [www.sourceforge.net/projects/febrl/](http://www.sourceforge.net/projects/febrl/)) are two products that health services researchers have used frequently in recent years. The authors are not aware of any scholarly evaluations of these products.

### **Commercially available products**

Select commercially available products have been reviewed by Day, Winkler, and Jones and Sujansky (Day, 1995; William, 2001; Jones and Sujansky, 2004). Here, we list three currently available products.

**LinkageWiz**, developed by LinkageWiz Software, is available at <http://www.linkagewiz.net/>. The cost of LinkageWiz ranges from \$199 for a 50,000 record limit to \$2,999 for an unlimited record limit.

**G-Link**, developed by Statistics Canada, is available at <http://www1.unece.org/stat/platform/display/msis/G-Link> for \$12,500 CAD. This product is based on the probabilistic theory developed by William Winkler.

**LinkSolv**, developed by StrategicMatching, is available at <http://www.strategicmatching.com/downloads.html>.

## **Chapter 5. An evaluation of methods linking health registry data to insurance claims in scenarios of varying available information**

### **Objective**

In this chapter, we expand upon our Chapter 4 discussion of linkage methods through an empirical linkage demonstration and evaluation using registry and insurance claims data. Here, we evaluate a set of linkage algorithms for registry-to-claims linkages covering scenarios of varying unique identifier availability, and incorporate encryption algorithms to allow linkage without PHI transfer. We evaluated test algorithms against a gold standard used by the National Cancer Institute's SEER-Medicare program. More specifically, we examined linkage algorithms first with full identifying information including name and Social Security Number among others, then through scenarios of iteratively fewer unique identifiers and greater reliance on non-unique information such as date of birth and sex. Given the exceptionally limited availability of practical empirical examples researchers can use to inform their own data linkages, this examination articulates much needed, specific details of the steps researchers may take and what they may expect to find given each study's unique scenario of data availability and quality.

### **Methods**

#### **Approach Overview**

We compared four approaches:

1. Employment of the current gold standard linking algorithm, first with full identifying information and subsequently with partial identifiers in place of their full counterparts
- 2a. Evaluation of deterministic approaches modeling scenarios of decreasing individually identifiable information
- 2b. Evaluation of deterministic approaches in the context of encrypted individual identifiers simulating a scenario of restrictions on identifier release to researchers
3. Evaluation of probabilistic approaches modeling scenarios of decreasing individually identifiable information

Experimental linkage sets will start with full available information and iteratively reduce the available information, in the end simulating a scenario in which unique identifiers are not available. To both streamline this examination and test the robustness of the algorithms in the context of the rareness of the condition of interest, we will begin by focusing on a smaller sample (subpopulation) with colon cancer, a common sex-neutral cancer. Next, we will examine cancers that are rarer and sex-specific. Finally, we will evaluate algorithm performance in the context of all cancers simultaneously in the full health plan claims population. Table 7 provides an overview of our approach.

**Table 7. Overview of experimental linkage approach**

	Maximum available information and unique Individual Identifiers (Gold Standard)	→→→	Incrementally reduced information, no unique individual identifiers (Experimental Sets)		
<b>1. Deterministic Linkage</b>	Linkage 1.1	Linkage 1.2	..	..	Linkage 1.n
<b>2. Deterministic with encryption</b>	Linkage 2.1	Linkage 2.2	..	..	Linkage 2.n
<b>3. Probabilistic Linkage</b>	Linkage 3.1	Linkage 3.2	..	..	Linkage 3.n

### Data Sources and patient populations

*Case data:* Individuals in the North Carolina Central Cancer Registry (NCCCR) diagnosed with colon cancer between years 2007-2008 (n = 6,444 unique individuals)

*Claims data:* Enrollment and claims data for 100% of North Carolina residents enrolled in any Blue Cross and Blue Shield (BCBS) owned- or administered plan, or the State Employees Health Plan (SEHP) for years 2006-2009 (n = 3,747,250 unique beneficiaries)

We selected these datasets because they have full information (i.e., all identifying variables) commonly captured in constituent datasets for linkages of this nature. The variables in the claims data are the same as those available in the Federal payer/claims data. Registry records will be matched to all years of BCBS/SEHP claims. The BCBS/SEHP claims data can be restricted to simulate practical scenarios of comparatively limited linking information experienced by researchers.

Table 8 shows identifiers available in both datasets.

**Table 8. Variables available for linkage and their completeness in study datasets**

Variable	Alternative forms	Completeness	
		CCR (%)	BCBS (%)
SSN	SSN4; SSN2	99.7	89.3
Last name	Last Name Initial (LNI); Last Name Soundex (LNS)	100.0	100.0
First name	First Name Initial (FNI); First name Soundex (FNS)	100.0	99.9
Date of birth	Month and year of birth; year of birth; DOB2	100.0	100.0
Sex		100.0	100.0
Residence	County of residence; Zip Code of Residence	99.9	99.9
Diagnosis	Valid Diagnosis code		100.0
	Diagnosis category (e.g., any cancer)	100.0	28.1
	Specific diagnosis (e.g., colon cancer)		2.9

Note: SSN4: Last 4 digits of SSN. SSN2: Last 2 digits of SSN. DOB2: 2 of 3 DOB parts.

### Data cleaning and standardization

Before the linkage, variables were cleaned and standardized as follows:

1. All string variables were converted to uppercase, stripped of all punctuation and digits, and hyphenated names were broken out into two different name fields.
2. All date variables were converted to date9. format (e.g., 01SEP2013).
3. Zip codes were limited to the first 5 digits.
4. FIPS codes were broken out into state (first 2 digits) and county (last 3 digits) codes.

5. Invalid SSNs (as defined by the Social Security Administration [http://ssa-custhelp.ssa.gov/app/answers/detail/a\\_id/425/~/~determining-social-security-numbers](http://ssa-custhelp.ssa.gov/app/answers/detail/a_id/425/~/~determining-social-security-numbers)) were flagged and treated as missing.

## **Data linkage**

### ***Blocking phase***

Rather than consider the Cartesian product of all possible matches between NCCCR and BCBS/SEHP, a subset of potential matches was identified during an initial blocking phase. Two records were included in the subset of potential matches if they agreed on any of the following:

1. SSN
2. Date of Birth, First Name Initial, and Sex
3. Date of Birth, Last Name Initial, and Sex
4. Last Name, First Name, and Sex
5. Date of Birth, County, and Sex

The blocking phase identified 104,360 possible matches.

### ***Step 1. Application of a Gold Standard Algorithm***

At this time, because there is presently no definitive gold standard algorithm for registry linkages, we used the linkage algorithm developed by the National Cancer Institute's Surveillance Epidemiology and End-Results (SEER)-Medicare program as a gold standard.(Warren, Klabunde et al. 2002) The iterative deterministic approach employed in this algorithm has demonstrated high validity and reliability in previous registry-to-claims linkages, has been employed successfully in numerous updates of the SEER-Medicare linked dataset, and is generally perceived to be strong in scenarios of high data quality and identifier completeness.(Potosky, Riley et al. 1993; Warren, Klabunde et al. 2002; National Cancer Institute SEER-Medicare Program 2011)

Individuals in the NCCCR database were linked to beneficiaries in the BCBS/SEHP database using the SEER-Medicare algorithm, which consists of a sequence of deterministic matches using different match criteria in each successive round:

- In the first step, records were declared a match if they agreed on SSN and one of the following:
- i. first and last name (allowing for fuzzy matches, such as nicknames)
  - ii. last name, month of birth, and sex, or
  - iii. first name, month of birth, and sex

If SSN was missing or did not match, or two records failed to meet the initial match criteria, they were subjected to a second round of deterministic linkages. In the second round, records were declared a match if they agreed on last name, first name, month of birth, sex, and one of the following:

- i. 7-9 digits of the SSN; or
- ii. Two or more of year of birth, day of birth, middle initial, or date of death

For each pair of records, match or non-match status was determined using the rules above, and match markers were generated indicating agreement or disagreement on each individual identifier. The SEER-Medicare algorithm classified 1,189 record pairs as matches, and 103,171 record pairs as non-matches. Based on prior knowledge of cancer incidence and insurance coverage in North Carolina, we expected that approximately 20% of individuals in the NCCCR database with colon cancer would be insured by BCBS/SEHP. The 1,189 uniquely matched individuals represent 18.5% of the individuals with colon cancer identified in the NCCCR database.

All subsequent test algorithms were evaluated against the SEER-Medicare algorithm. The match decisions made by each algorithm were compared with the match decisions made through use of the gold-standard algorithm. Pairs identified as matches by both the SEER-Medicare algorithm and the test algorithm were declared to be “true matches”. Pairs identified as matches by the SEER-Medicare algorithm and non-matches by the test algorithm, were declared to be “false non-matches”. Pairs identified as non-matches by both the SEER-Medicare algorithm and the test algorithm were declared to be “true non-matches”. Pairs identified as non-matches by the SEER-Medicare algorithm and matches by the test algorithm were declared to be “false matches”.

To assess the success of each algorithm, we calculated sensitivity, positive predictive value, and f-measure using SAS (version 9.3; SAS Institute, Cary, NC).

### ***Step 2a. Comparison Approach 1 – Deterministic Linking***

Deterministic linkage strategies have been recommended for situations in which the data are of a high quality and/or many identifier variables are available.(Wajda and Roos 1987) Research has also shown that deterministic linkage on a sufficient number of partial and/or indirect identifiers such as initials, year of birth, and county of residence, can provide sufficient discriminatory power to classify matches and non-matches with good sensitivity and specificity.(Roos and Wajda 1991; Jamieson, Roberts et al. 1995)

Using the matching markers generated above, we developed and tested a set of deterministic algorithms, using match variable combinations of full and partial identifiers, covering information-rich situations in which full direct identifiers (e.g., SSNs and full names) are available, as well as information-poor situations in which only indirect or partial identifiers are available. Only record pairs that were uniquely identified by the given variable combination were considered as potential matches. When multiple record pairs matched on the values of a given variable combination, the record pairs were flagged as ties and classified as non-matches. To account for minor typographical errors in names, we used the Soundex algorithm to generate a code consisting of the first initial and up to three digits representing consonant sounds in the name, and matched on the Soundex values. By doing so, we were able to explore the possibility of linking algorithms that do not require the release of full or actual names.

Following Roos and Wajda, we determined the percentage of records identified uniquely by each combination of variables (Wajda and Roos 1987). Given the exploratory nature of this study, we relaxed the recommended threshold of  $\sim 1.00$  records per unique value (100% uniqueness), and included for testing all variable combinations that identified 85% of records uniquely. Using this method, we selected 398 variable combinations for testing. To simulate scenarios of decreasing information availability, the variables with the largest number of unique values were removed in a stepwise manner. The first group of algorithms used all identifiers. The second group of algorithms excluded SSN. Finally, the third group of identifiers excluded SSN and name. The five best performing algorithms in each group are reported in Table 3.

### ***Step 2b. Comparison Approach 1, encryption variation***

In situations where full identifiers or partial identifiers are available but may not be released or transmitted, research has shown that records can be successfully linked via deterministic algorithms using identifiers encrypted before release. (Quantin, Bouzelat et al. 1998; Quantin, Bouzelat et al. 1998; Quantin, Allaert et al. 2008) To simulate the application of a hash encryption method before release, we converted the variable combinations presented in Step 2a to 128-bit hash values using the md5 algorithm. Each conversion was performed using the md5 function in SAS 9.3. It is important to note that the length, format, order, and content of the strings in the two datasets have to be perfectly consistent before the conversion. If there is even a slight difference between the two strings, the md5 algorithm will generate two different values, as shown in Table 9 below.

**Table 9. Example md5 algorithm values from inconsistent strings**

---

SOURCE	DATE OF BIRTH	FIRST NAME	LAST NAME	MD5 CODE
NCCCR	12312013	BILL	SMITH	<DgiÖ oÓli=\$2e'Ä
BCBS/SEHP	12312013	Bill	Smith	*Ym1" Ö¼âëÑöÁ@

While the two records in this example clearly match on date of birth, first name, and last name, the md5 hash values for the two concatenated strings are very different due to the different casing on the names. Both strings would need to be ordered, formatted, and spaced uniformly for the md5 algorithm to generate the same value for the two strings. Using the example above, the best approach would be to concatenate the three identifiers, remove all spaces, and convert the case of the names to uppercase (ie. '12312013BILLSMITH') before applying the md5 algorithm. We performed a deterministic match on the hash values for each variable combination presented in Step 2a.

### ***Step 3. Comparison Approach 2 – Probabilistic Linking***

Probabilistic linkage strategies have been recommended for situations in which the data contain many coding errors and/or only a few identifiers are available (Wajda and Roos 1987). Using the match markers generated earlier, we developed and tested a set of probabilistic algorithms using the match variable combinations in each group of full and partial identifiers that performed best in Step 2a. We covered information-rich situations in which full direct identifiers (e.g., SSNs and full names) are available, as well as information-poor situations in which only indirect or partial identifiers are available. Only record pairs that were uniquely identified by the given variable combination were considered as potential matches. When multiple record pairs matched on the values of a given variable combination, the record pairs were flagged as ties and classified as non-matches. The goal in this step is to improve on the match results in Step 2a by making use of the information ignored in deterministic algorithms.

For each matched pair, we calculated agreement weights and disagreement weights for each identifier. Following the Fellegi and Sunter model,(Fellegi and Sunter 1969) agreement weights were calculated by dividing the probability that true matches agree on the specific value of the identifier by the probability that false matches randomly agree on the specific value of the identifier, and taking the  $\log_2$  of the quotient. For example, if the probability that true matches agree on month of birth is 97% and the probability that false matches randomly agree on month of birth is 8.3% (1/12), then the agreement

weight for month of birth would be  $\log_2(.97/.083)$ , or, 3.54. Disagreement weights were calculated by dividing 1 minus the probability that true matches agree on the specific value of the identifier by 1 minus the probability that false matches agree on the specific value of the identifier, and taking the  $\log_2$  of the quotient.

To allow for comparisons across linkage strategies, we selected for testing the same 398 variable combinations that that were selected for testing in Step 2a. The linkage score for each matched pair was then computed as the sum of the weights. Using the method developed by Cook et al., (Cook, Olson et al. 2001) we calculated the threshold weight needed to achieve a 95% probability that two matched records are a true match. Matched pairs with a linkage score greater than the threshold weight were declared “matches”, while matched pairs with a linkage score less than the threshold weight were declared “non-matches”. The top five algorithms are presented in Table 13.

## Results

### *Gold Standard Linkage*

The SEER-Medicare algorithm, using full identifiers, classified 1,189 record pairs as matches, and 103,171 record pairs as non-matches. In a stepwise fashion, we replaced the full identifiers with partial identifiers to determine whether the algorithm can work in the absence of full identifiers. Selected results of the SEER-Medicare iterative deterministic algorithm with full identifiers replaced with the indicated partial identifiers are presented in Table 10. The results indicate that the sensitivity of the algorithm was unaffected in the three examples presented in the table. The replacement of full identifiers with partial identifiers, however, did slightly increase the number of true non-matches classified as matches (Note: manual review confirmed that the additional matches identified by the algorithms with partial identifiers were in fact non-matches). Despite the small decrease in the specificity of the algorithm, these results indicate that the SEER-Medicare linkage can perform effectively in the absence of full identifiers.

**Table 10. Select results of gold standard linkage algorithm with partial identifiers**

Linking Variables	Matches		Non-Matches		Sensitivity	PPV	F-Measure
	True	False	True	False			
Algorithm1	1,186	14	103,157	3	99.75	98.83	99.29
Algorithm2	1,185	13	103,158	4	99.66	98.91	99.28

Algorithm3	1,189	19	103,152	0	100.00	98.43	99.21
Algorithm4	1,171	9	103,162	18	98.49	99.24	98.86
Algorithm5	1,171	14	103,157	18	98.49	98.82	98.65

Note: LNS: Last Name Soundex. FNS: First Name Soundex. SSN: Social Security Number. SSN4: Last 4 digits of SSN. SSN2: Last 2 digits of SSN. DOB2: 2 of 3 DOB parts. MI: Middle Initial

Algorithms are provided in detail in Appendix 4.

### ***Deterministic Linkage***

Results of the deterministic linkages are presented in Table 11 on the next page. The relatively lower sensitivity scores (87.13-88.39) for algorithms using SSN reflect the fact that only 89% of BCBS members had a valid SSN. As expected, algorithms using SSN have very high specificity (99.99-100.00) and positive predictive value (99.33-99.90).

When we excluded SSN, the best performing algorithms were able to identify correctly more matches (85.70-92.26) without sacrificing specificity (99.99-100.00), and only minor decreases in positive predictive value (99.03-100.00). The most encouraging result is the finding that Date of Birth, Last Name Soundex, First Name Soundex, and Sex correctly uniquely identified 92% of matches identified by the SEER-Medicare algorithm, with specificity and positive predictive value over 99%. Preferably, all values would be greater than 95%, but this finding demonstrates that a good linkage can be performed in the absence of SSN or actual name.

**Table 11. Select results of deterministic linkage algorithms**

Linking Variables	Matches		Non-Matches		Sensitivity	PPV	F-Measure
	True	False	True	False			
<b>Combinations that include SSN</b>							
SSN4 and Month of Birth	1,051	6	103,165	138	88.39	99.43	93.59
SSN4 and Year of Birth	1,048	4	103,167	141	88.14	99.62	93.53
SSN and Month of Birth	1,041	3	103,168	148	87.55	99.71	93.24
SSN and Sex	1,043	7	103,164	146	87.72	99.33	93.16
SSN and Year of Birth	1,036	1	103,170	153	87.13	99.90	93.08
<b>Excluding SSN</b>							
DOB, FNS, LNS, and Sex	1,097	5	103,166	92	92.26	99.55	95.77
DOB, FN, LN, and Sex	1,087	0	103,171	102	91.42	100.00	95.52
DOB2, FNS, LNS, County, and Sex	1,029	9	103,168	160	86.54	99.13	92.41
DOB, LN, County, and Sex	1,020	5	103,166	169	86.12	99.51	92.33

DOB, LNS, County, and Sex	1,019	10	103,161	170	85.70	99.03	91.88
<b>Excluding SSN and Name</b>							
DX, DOB2, Zip, and Sex	839	12	103,159	350	70.56	99.59	82.60
DX, DOB, County, and Sex	841	29	103,142	348	70.73	96.67	81.69
DX, DOB, Zip, and Sex	824	9	103,162	365	69.30	98.92	81.50
DX, Year of birth, Zip, and Sex	813	9	103,162	376	68.38	98.19	80.62
DX, Month of birth, Zip, and Sex	749	8	103,163	440	62.99	98.94	76.97
<b>Excluding SSN, Name and DOB</b>							
DX, Zip, MI, and Sex	552	3	103,168	637	46.43	99.46	63.31
DX, Zip, and MI	541	3	103,168	648	45.50	99.45	62.43
DX, County, MI, and Sex	394	4	103,167	795	33.14	98.99	49.66
DX, County, and MI	333	3	103,168	856	28.01	99.11	43.68
DX, Zip, and Sex	332	4	103,167	857	27.92	98.91	43.55

Note: LNS: Last Name Soundex. FNS: First Name Soundex. SSN: Social Security Number. SSN4: Last 4 digits of SSN. SSN2: Last 2 digits of SSN. DOB2: 2 of 3 DOB parts. MI: Middle Initial

The sensitivity of algorithms that did not include SSN or name was significantly lower than algorithms that did include SSN and/or name. Once ties were classified as non-matches, the sensitivity of these algorithms was 70.23 – 73.25. However, algorithms that blocked on primary site (e.g., diagnosis code for colon cancer), demonstrated high specificity (99.96-99.99) and high positive predictive value (95.09-99.59).

### ***Deterministic linkage with encryption***

Results of the deterministic linkage approaches using encryption are presented in Table 12 on page 78. The results for each algorithm were consistent with the previous results (Table 11), indicating that a deterministic match on identifiers encrypted before release can be successful in instances where identifiers are available, but not releasable.

### ***Probabilistic Linkage***

As shown in Table 13 on page 79, the probabilistic approach improved the performance of all algorithms. When all identifiers were included, the sensitivity improved from ~87% to 97.92%, because many of the ~13% of BCBS members missing SSN were matched using information provided by matches on other identifiers.

This demonstrates the ability of probabilistic algorithms to perform when data quality is poor. In this instance, missing information in one important identifier was overcome by information provided in other identifiers, thus improving the sensitivity and accuracy of the probabilistic approach compared with a deterministic approach. While the iterative deterministic approach used in the SEER-Medicare algorithm is similarly able to overcome poor data quality in an important identifier such as SSN, it relies on SSN and full name. Conversely, probabilistic algorithms can be effective in scenarios where SSN and full name are unavailable, as demonstrated by the second probabilistic algorithm reported in Table 4. Using only date of birth, first and last name Soundex values, residence, diagnosis, and sex, the probabilistic approach was still able to identify correctly 96.67% of true matches and 99.99% of true non-matches. Thus, if confidentiality concerns block the release of SSN and full name in the future, registry data can still be linked successfully to claims using the probabilistic approach. The final result reported in Table 4 shows that algorithms relying solely on date of birth, residence, diagnosis, and sex were unsuccessful, though the probabilistic approach showed some improvement over the deterministic approach. Additional information not used in this study (e.g., service dates) may provide a probabilistic approach with the added power needed for a successful linkage.

**Table 12. Select results of deterministic linkage algorithms using encrypted data**

Linking Variables	Matches		Non-Matches		Sensitivity	PPV	F-Measure
	True	False	True	False			
<b>Combinations that include SSN</b>							
SSN4 and Month of Birth	1,051	6	103,165	138	88.39	99.43	93.59
SSN4 and Year of Birth	1,048	4	103,167	141	88.14	99.62	93.53
SSN and Month of Birth	1,041	3	103,168	148	87.55	99.71	93.24
SSN and Sex	1,043	7	103,164	146	87.72	99.33	93.16
SSN and Year of Birth	1,036	1	103,170	153	87.13	99.90	93.08
<b>Excluding SSN</b>							
DOB, FNS, LNS, and Sex	1,097	5	103,166	92	92.26	99.55	95.77
DOB, FN, LN, and Sex	1,087	0	103,171	102	91.42	100.00	95.52
DOB2, FNS, LNS, County, and Sex	1,029	9	103,168	160	86.54	99.13	92.41
DOB, LN, County, and Sex	1,020	5	103,166	169	86.12	99.51	92.33
DOB, LNS, County, and Sex	1,019	10	103,161	170	85.70	99.03	91.88
<b>Excluding SSN and Name</b>							
DX, DOB2, Zip, and Sex	839	12	103,159	350	70.56	99.59	82.60
DX, DOB, County, and Sex	841	29	103,142	348	70.73	96.67	81.69
DX, DOB, Zip, and Sex	824	9	103,162	365	69.30	98.92	81.50

DX, Year of birth, Zip, and Sex	813	9	103,162	376	68.38	98.19	80.62
DX, Month of birth, Zip, and Sex	749	8	103,163	440	62.99	98.94	76.97
<b>Excluding SSN, Name and DOB</b>							
DX, Zip, MI, and Sex	552	3	103,168	637	46.43	99.46	63.31
DX, Zip, and MI	541	3	103,168	648	45.50	99.45	62.43
DX, County, MI, and Sex	394	4	103,167	795	33.14	98.99	49.66
DX, County, and MI	333	3	103,168	856	28.01	99.11	43.68
DX, Zip, and Sex	332	4	103,167	857	27.92	98.91	43.55

Note: LNS: Last Name Soundex. FNS: First Name Soundex. SSN: Social Security Number. SSN4: Last 4 digits of SSN. SSN2: Last 2 digits of SSN. DOB2: 2 of 3 DOB parts. MI: Middle Initial.

**Table 13. Select results of probabilistic linkage algorithms:**

Linking Variables	Matches		Non-Matches		Sensitivity	PPV	F-Measure
	True	False	True	False			
<b>Combinations that include SSN</b>							
SSN4, FN, LN, DOB, County, and Sex	1,171	7	103,164	18	98.49	99.41	98.95
SSN4, FN, LN, DOB, Zip, and Sex	1,171	8	103,163	18	98.48	99.32	98.90
SSN, FN, LN, DOB, Zip, and Sex	1,171	8	103,163	18	98.48	99.32	98.90
SSN, FNS, LNS, DOB, Zip, and Sex	1,169	8	103,163	20	98.32	99.32	98.82
SSN4, FNS, LNS, DOB, Zip, and Sex	1,168	8	103,163	21	98.23	99.32	98.77
<b>Excluding SSN</b>							
DOB, LN, FN, Zip, and Sex	1,147	10	103,161	42	96.47	99.14	97.79
DOB, LN, FN, County, and Sex	1,136	9	103,162	53	95.54	99.21	97.34
DOB, LNS, FNS, County, and Sex	1,119	9	103,162	70	94.11	99.93	96.93
DOB, LNS, FNS, Zip, and Sex	1,131	14	103,157	58	95.12	98.78	96.92
DOB, LN, Zip, and Sex	1,033	10	103,161	156	86.88	99.04	92.56
<b>Excluding SSN and Name</b>							
DX, DOB, Zip, MI, and Sex	885	22	103,149	304	74.43	97.57	84.44

DX, DOB2, Zip, MI, and Sex	865	17	103,154	324	72.75	98.07	83.53
DX, DOB, Zip, and Sex	830	9	103,162	359	69.81	98.93	81.86
DX, DOB, County, and Sex	844	29	103,142	345	70.98	96.68	81.86
DX, Year of birth, Zip, and Sex	818	9	103,162	371	68.71	98.91	81.09
<b>Excluding SSN, Name and DOB</b>							
DX, Zip, MI, and Sex	765	10	103,161	424	64.34	98.71	77.90
DX, Zip, and MI	719	7	103,164	470	60.47	99.04	75.09
DX, County, MI, and Sex	634	9	103,162	555	53.12	98.60	69.04
DX, County, and MI	333	3	103,168	856	28.01	99.11	43.68
DX, Zip, and Sex	333	4	103,167	856	28.01	98.81	43.65

Note: LNS: Last Name Soundex. FNS: First Name Soundex. SSN: Social Security Number. SSN4: Last 4 digits of SSN. SSN2: Last 2 digits of SSN. DOB2: 2 of 3 DOB parts. MI: Middle Initial

## Discussion

The results of this study indicate that a successful linkage is possible in the absence of full identifying information. We found that straightforward and easy-to-employ deterministic algorithms using date of birth and Soundex codes for names demonstrated high specificity and positive predictive value with acceptable sensitivity. In situations where identifiers are available, but not allowed to be released, we found that deterministic matching on hash-encrypted variable combinations performed equally well as deterministic matching on the same combination of unencrypted variables.

In information-rich scenarios where identifiers are available for release, iterative deterministic approaches such as the SEER-Medicare algorithm are highly effective, and much more time and resource efficient compared with probabilistic approaches, which can be highly complex and difficult to implement. However, when unique identifiers such as SSN and full name are unavailable, the probabilistic approach consistently outperforms the deterministic approach. These findings are particularly important as confidentiality concerns are making it increasingly difficult to obtain identifying information for linkage projects.

## References

1. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Medical care*. Aug 2002;40(8 Suppl):IV-3-18.
2. National Cancer Institute SEER-Medicare Program. Search SEER-Medicare Publications. 2011; <http://healthservices.cancer.gov/seermedicare/overview/publications.html>. Accessed March 4, 2011.
3. Potosky AL, Riley GF, Lubitz JD, Mentnech RM, Kessler LG. Potential for cancer related health services research using a linked medicare-tumor registry database. *Medical Care*. 1993;31(8):732-748.
4. Wajda A, Roos LL. Simplifying record linkage: software and strategy. *Computers in biology and medicine*. 1987;17(4):239-248.
5. Jamieson E, Roberts J, Browne G. The feasibility and accuracy of anonymized record linkage to estimate shared clientele among three health and social service agencies. *Methods of Information in Medicine*. 1995;34:371-377.
6. Roos LL, Wajda A. Record linkage strategies. Part 1: estimating information and evaluating approaches. *Methods of Information in Medicine*. 1991;30:117-123.
7. Quantin C, Allaert F-A, Avillach P, et al. Building application-related patient identifiers: What solution for a European Country? *International Journal of Telemedicine and Applications*. 2008:1-5.
8. Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow up: quality assessment of an anonymous record linkage procedure. *International Journal of Medical Informatics*. 1998;49:117-122.
9. Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods of Information in Medicine*. 1998;37(3):271-277.
10. Fellegi IP, Sunter AB. A Theory for Record Linkage. *Journal of the American Statistical Association*. 1969;64(328):1183-1210.
11. Cook LJ, Olson LM, Dean JM. Probabilistic Record Linkage: Relationships between File Sizes, Identifiers, and Match Weights. *Methods of Information in Medicine*. 2001;40:196-2003.

## Appendix 4. SEER-Medicare algorithm with partial identifiers

### *Algorithm 1:*

```

if last 4 digits of ssn match then
    if first name soundex, last name soundex, 2 out of 3 dob parts match
    or
    if last name soundex, 2 out of 3 dob parts, sex match
    or
    if first name soundex, 2 out of 3 dob parts, sex match
    then it's a match

if last 4 digits of ssn do not match then
    if last name soundex, first name soundex, 2 out of 3 dob parts, sex match then
        if sum(of middle initial, date of death) >= 1
        or
        if sum(of zip, county) >= 1
        or
        if primary_site match
    then it's a match
  
```

### *Algorithm2:*

```

if last 4 digits of ssn match then
    if first name soundex, last name soundex, 2 out of 3 dob parts match
    or
    if last name soundex, 2 out of 3 dob parts, sex match
    or
    if first name soundex, 2 out of 3 dob parts, sex match
    then it's a match

if last 4 digits of ssn do not match then
    if last name soundex, first name soundex, 2 out of 3 dob parts, sex then
        if sum(of middle initial, date of death) >= 1
        or
        county
    then it's a match
  
```

### *Algorithm3:*

```

if last 4 digits of ssn match then
    if first name soundex, last name soundex match
    or
    last name soundex, 2 of 3 date of birth parts, sex match
    or
    first name soundex, 2 out of 3 dob parts, sex match
    then it's a match
  
```

if last 4 digits of ssn do not match then  
 if last name soundex, first name soundex, month of birth, sex match  
 or  
 if (sum(of year of birth, day of birth, middle initial, date of death)  $\geq$  2)  
 then it's a match

***Algorithm 4:***

if last 4 digits of ssn match then  
 if first name soundex, last name soundex, 2 out of 3 dob parts match  
 or  
 if last name soundex, 2 out of 3 dob parts, sex match  
 or  
 if first name soundex, 2 out of 3 dob parts, sex match  
 then it's a match

if last 4 digits of ssn do not match then  
 if last name soundex, first name soundex, date of birth, sex match  
 or  
 if last name soundex, 2 out 3 date of birth parts, zipcode, sex, (middle initial or date of death)  
 match  
 or  
 if first name soundex, 2 out 3 date of birth parts, zipcode, sex, (middle initial or date of death)  
 match  
 then it's a match

***Algorithm 5:***

if last 4 digits of ssn match then  
 if first name soundex, last name soundex, 2 out of 3 dob parts match  
 or  
 if last name soundex, 2 out of 3 dob parts, sex match  
 or  
 if first name soundex, 2 out of 3 dob parts, sex match  
 then it's a match

if last 4 digits of ssn match then  
 if last name soundex, first name soundex, date of birth, sex match  
 or  
 if last name soundex, 2 out 3 date of birth parts, zipcode, sex match  
 then it's a match

## Chapter 6. Project summary and recommendations for researchers

### Overview

Randomized controlled trials (RCTs) remain the gold standard for assessing intervention efficacy; however, RCT results often cannot be generalized due to a lack of inclusion of “real world” combinations of interventions and heterogeneous patients (Clancy and Slutsky 2007; IOM 2009; Congressional Budget Office 2007; Smith, 2007). With recent advances in information technology, data, and statistical methods, there is tremendous promise in leveraging ever-growing repositories of secondary data to support comparative effectiveness and public health research (Smith 2007; IOM 2009; Sox 2006). While there are many advantages to using these secondary data sources, they are often limited in scope, which, in turn, limits their utility in addressing important questions in a comprehensive manner. These limitations can be overcome by linking data from multiple sources such as health registries and administrative claims data (VanLare, Conway, et.al. 2010; Bloomrosen, 2008; CDC 2010; Sturmer 2011).

This report provides a conceptual framework for high-quality data linkage in the context of comparative effectiveness research (CER). It describes the infrastructure and personnel needed to navigate the linkage process, outlines the DUA process, presents different approaches to data linkage, discusses the strengths and weaknesses of each, and makes recommendations on which approaches perform best in varying scenarios of data availability. Through this report, researchers have a step-by-step instructional guide for designing new CER studies that involve linking patient registries with other secondary data sources. We also highlight considerations for researchers, data managers, information technology managers, and other stakeholders likely to be involved in the data linkage process.

### Considerations for project planning

#### Appropriateness and feasibility of the project

Each data source needs to be considered carefully regarding its adequacy for the specific linkage endeavor. First and foremost, the *quality* and *discriminatory power* of each of the available linkage identifiers needs to be scrutinized. Second, researchers must be confident that there is enough overlap in the two populations to merit the effort. Every linkage will result in a set of matches that will serve as a select, though perhaps not completely representative (or random) sample, of the two disparate underlying populations. It is important that the extent and representativeness of this potential overlap be

considered before beginning the process. Finally, it must be evident that the linkage will result in an enriched dataset that includes additional data elements made possible only through the linkage. Each of these considerations must be weighed together as a whole to determine whether the linkage will provide an adequate return on the investment in time and resources.

### **Data ownership and governance**

Before proposing a linkage, it must be clear that each of the respective data sources allows for the scope of the proposed work. All data owners or key stakeholders need to be contacted and existing data governance processes need to be understood. Rules regarding consent and any existing regulatory requirements for the data need to be clarified. The importance of these issues cannot be underestimated and the logistical, administrative, and often legal hurdles need to be anticipated and built into the cost and the timeline for the project.

### **Technical environment and security**

A secure and well-performing computing platform represents the operational backbone for conducting innovative research using large datasets. There is a fine balance between security and usability, and system performance directly influences the size and complexity of the data that can be managed and linked. Careful planning and building collaborations with teams with technical resources helps control for the cost/benefit factor while allowing growth in the future. As outlined in *Chapter 2*, a well-performing and secure research environment builds a baseline for trust with research partners and data suppliers.

### **Team, skills, and expertise**

The complexity and scope of a research project will dictate the type of team and expertise required. As scope and complexity increase, the required experience and expertise of team members narrows and deepens. Ideally, a research and data support team will already be in place before the proposal process. These individuals can provide essential insight into considerations such as feasibility, technical environment, approach, and linkage processes identified in this report. Specific and key roles have been outlined in *Chapter 2*. For nearly all data linkage projects, the research team should include individuals with experience developing and evaluating linking algorithms, as well as overall expertise in population research. Other technical skills are no less essential for project success. For example, a knowledgeable data manager will be able to evaluate the appropriateness of the datasets, assess the computational feasibility of the linkage, estimate computing requirements, and develop a common data model.

Information or computing technology experts can help design, run, and maximize the required technical environment and computing platform. Lastly, any large linkage project will require a conscientious and knowledgeable project manager who is experienced with the many security standards and legal documents and processes that are required to get the project off of the ground and keep it moving throughout its life course.

### **Cost**

Several financial considerations need to be incorporated into project planning and execution for any linkage project. First, obtaining a copy of the data or purchasing a user license often comes with a hefty price. Second, the IT systems and technical platform requirements are often significant and can be expensive to build and maintain. However, other options may be available besides building a *de novo* system. Renting infrastructure from another research partner or computing environment, or utilizing a third-party vendor for data linkage may provide cost-savings and synergy at an organizational level. Lastly, the technical personnel required for linking and maintaining data are often in high demand in this era of “big data” and may require high salaries to recruit and retain.

### **Identify and evaluate available linkage keys**

The quantity, quality, and discriminatory power of available identifiers will determine the feasibility and success of any linking endeavor. This information may or may not be known before data delivery, but can often be estimated. *Chapter 3* outlines different types of identifiers, key aspects of each, and identifies factors to consider when weighing the feasibility of a proposed linkage project. The quality, completeness, and predictive ability of each of the potential keys must be assessed separately in each respective dataset, with careful attention paid to missing, invalid, or implausible values. This effort will drive the choice of the appropriate data linkage strategy.

### **Variable cleaning, standardization, and a common data model (Normalization)**

To begin cleaning and standardizing variables before linking, it is vital that the researchers obtain and review all available reference documentation (e.g., literature search, grey literature), so the research team understands the underlying intent and purpose of the data, as well as the sources or origination of the variables. These aspects of the original data directly affect the quality of the variables and drive later decisions about the appropriate linkage method. The amount of data cleaning and standardization needed depends on the quality and source of the data as well as the researcher’s question and available resources. *Chapter 4* of this report outlines some of the common standardization measures required to

prepare the linkage identifiers. Each key variable comes with its own idiosyncrasies, and thus a specific standardization and cleaning protocol must be developed and documented for each key. Specifically, decisions regarding standardization of format (e.g., dates set to ‘mmddyyyy’), case (e.g., all uppercase), punctuation (e.g., remove dashes from SSN), and missing values (e.g., use a consistent value: ‘’, ‘’, ‘9999’) need to be made and documented. In some cases secondary software or datasets (cross-walk files) may help to standardize or clean the variables (e.g., linkage to common nicknames or synonyms).

Downstream from standardizing and cleaning identifiers but important in preparing the data for linkage, is the development of a common data model. As the data are prepared for linkage and subsequent use, researchers need to understand, outline, define, and document the division of the data into separate sub-tables, as well as the relationships between tables within the linked dataset (i.e., entity relationships). This involves the development of a procedure for joining tables (e.g., assignment of a primary key) and determining whether the relationships of the individuals within these tables are one-to-one, one-to-many, many-to-one, or many-to-many. Finally, decisions regarding variables that are repeated across datasets need to be made. Elements such as sex, race, and dates are often stored in different formats and need to be standardized in the linked dataset. Furthermore, protocols for handling discrepancies (e.g., different birth dates or race categories) across data sources need to be developed and documented.

### **Linkage approach**

Once the researchers have familiarized themselves with the relevant data sources and evaluated the available linkage variables, they can make an initial determination about the most appropriate data linkage strategy. *Chapters 4 and 5* outline different linkage approaches and discuss the strengths and weaknesses of each approach. The first step in selecting the appropriate linkage strategy is to determine whether direct unique identifiers are available (e.g., SSN). In scenarios in which direct unique identifiers are available, deemed to be of high quality, and non-missing in approximately 95% of cases in each dataset, a deterministic approach is recommended. A one-time deterministic approach is the easiest to design, implement, and interpret. It involves a binary, “all or nothing” decision-making process in which record pairs are compared character-for-character across all identifiers. Record pairs that agree exactly on the given identifiers are classified as matches, while record pairs that disagree on even a single character are classified as non-matches. This approach does not account for the differential *discriminatory power* of the identifiers or the *degree* to which record pairs agree. Deterministic approaches typically have high positive predictive value, but often suffer from low sensitivity due to the inflexibility of the criteria.

An iterative deterministic approach, such as the well-documented SEER-Medicare algorithm, provides a more flexible alternative to a one-time deterministic approach. It involves an initial match on the most conservative matching criteria, followed by subsequent matches where record pairs that failed to meet the initial criteria are passed to a second, more lenient set of matching criteria. Record pairs that meet the matching criteria on any step are classified as matches, while record pairs that fail each step of matching criteria are classified as non-matches. This approach is more flexible and more sensitive than the one-time deterministic approach. In scenarios in which no single unique identifier (e.g., beneficiary id or SSN) is available or complete, but high-quality direct identifiers are available (e.g., names, dates of birth, etc.), the SEER-Medicare algorithm has demonstrated high reliability and validity.

In many cases, identifiers are available but incomplete, fraught with typographical errors, or imperfectly measured. In these scenarios, probabilistic techniques are recommended, as they have consistently outperformed deterministic techniques in earlier research. As described in *Chapter 4*, a probabilistic approach incorporates the differential discriminatory power of the identifiers and the degree to which two records agree into agreement and disagreement weights for each identifier. This approach can be used upfront to initially test and validate the discriminatory power of each available identifier. During the linking process, probabilistic methods assign an agreement weight when identifiers agree or a disagreement weight when identifiers disagree, and then derive an overall match score based on the sum of all weights. While probabilistic methods often outperform deterministic methods in information-poor scenarios, they require significantly more time, effort, and technical resources to implement.

An optimal approach that covers all scenarios, datasets, research questions, and/or situations does not exist. Often, as outlined in *Chapter 5*, combining probabilistic and deterministic methods can be more efficient and save computational resources. For example, a deterministic match on all direct identifiers can be executed first to identify certain matches and the remaining discordant pairs evaluated using probabilistic matching. The decision of which approach to use depends ultimately on the research question and the available resources.

### **Evaluation and validation of record linkage**

Irrespective of the linkage method applied, careful evaluation of the output is needed. This begins with a manual review process. Review involves several steps, the first of which involves looking for and resolving ties in which multiple record pairs identified as matches by the algorithm have the exact same

set of values. Next, a random sample of the set of potential matches identified during the blocking phase should be reviewed to evaluate the accuracy of the algorithm and identify circumstances in which the algorithm can be refined to account for complicated cases or unforeseen patterns. Following the manual review process, the final proportion of matches should be compared to known metrics or expected (*a priori*) estimates. A substantial difference in the observed number of matches versus the expected number of matches is a good indication that the linkage approach needs adjustment.

### **Recommendations for reporting results**

Data linkage is a complex process that involves many decisions that may affect the validity and generalizability of the newly linked data source. We recommend that researchers who implement a data linkage to generate an analytic dataset report of the results of their data linkage process in manuscripts that utilize the linked data. Specifically, we recommend that researchers use the following statistical measures of performance: sensitivity, positive predictive values, and the f-measure to characterize the accuracy of the reported strategy. Specificity and negative predictive value, frequently reported in health services research, will be inflated because of the large number of true non-matches, and thus should not be reported. The metrics should be clearly presented for any algorithm so potential users can weigh their priorities (e.g., sensitivity over positive predictive value) and choose a strategy based on the needs of the specific project. These measures could be reported in the methods section of manuscripts that utilize the linked data.

Following the matching process, researchers should determine whether there are any important differences in the characteristics of people who are matched and those who are unmatched. The extent to which unmatched individuals differ from matched individuals should be reported within manuscripts using the linked data. This may be as simple as comparing characteristics of the new and original samples. This is particularly important when researchers link existing cohorts with external data sources. If the existing data source has been used extensively (for example, SEER Medicare), then identifying any differences between the original (known) cohort and the new cohort are critical to providing transparency to the reader regarding changes in the underlying population post-linkage.

## Framework for registry-to-claims linkage

We provide several checklists for researchers to use for registry-to-claims linkage project planning and execution. These are conceptualized as 1) project planning or what you need to consider before applying for funding, and 2) project execution or what steps you need to take in order to successfully execute the research project.

### Project planning checklist

Appropriateness and feasibility

Correct data for research question? Will I have linkage variables, overlap in population? Adequate Return on Investment (ROI)? Is there enough overlap between datasets? Are additional data elements/variables gained from linkage complete (non-missing) on the linked (overlapping) population? [effort to link = appropriate gains in data/population]?

Data ownership and governance

Do the data source allow for the scope of the proposed work? Are there existing data governance processes in place? If needed, did the subjects consent to the research?

Technical environment

Do I have plans to build or leverage a technical environment compliant with security needs? Do I have enough computing power? Do I have enough storage and tools to handle big data?

Team, skills, and expertise

Do I have access to a team or existing organizational entities for all aspects of the project? Data management; information confidentiality and security; linking expertise; epidemiologic expertise (re: selection bias, design, etc.)

Cost

Do I have estimates on the cost for advanced expertise? Are the costs for the technical platform feasible over the duration of the project? Do I build or “rent”? Can I leverage aspects of the project to attract follow-up funding?

### Project execution checklist

Build data partnerships and develop Data Use Agreements (DUA)

Identify stakeholders including necessary, legal, regulatory, or administrative staff. Identify regulatory/security requirements. Develop legal documents/agreements approved by all stakeholders.

Identify and recruit technical skills and expertise required

Find IT systems partner, data security officer, data manager, linkage programmer, epidemiologic/population expertise to help guide decision making

Identify and evaluate available identifiers/linkage keys

Determine the completeness and quality of variables needed for linkage. Evaluate missing, invalid, or unlikely values. Use this assessment to determine the appropriate data linkage strategy.

Standardize and clean linkage keys

Obtain and review ALL reference documentation available from data partners/vendors or published elsewhere (literature search, grey literature). Outline the standardization protocol for each identifier. Identify any secondary software or datasets (crosswalk files) you need to standardize or clean. Amount of cleaning required will depend on the quality and source of the data.

Developing a common data model / Normalization of data

Determine the entity relationships (one-to-many / one-to-one), standardize common variables across datasets (e.g., sex, race), standardize formatting (e.g., dates set to mmddyyyy), standardize case (e.g., all uppercase) and punctuation (e.g., remove dashes from SSN), determine how to handle missing values (e.g., use a consistent value: ‘’, ‘9999’) and duplicates.

Design linkage approach

By incorporating information from the previous steps, decide whether a deterministic linkage strategy is feasible by evaluating any direct identifiers for quality and completeness. Often probabilistic methods or a combined approach must be applied. Evaluate whether and how a “blocking strategy” can be used to improve efficiency and computing requirements. Be sure to incorporate necessary iterative processes including specific steps for review/evaluation.

Evaluate / validate record linkage and reporting of linkage metrics

Conduct a manual review of match decisions. The extent and sampling of this review may vary and will be defined by the defined linkage strategy. Compare results back with target or reference populations. Calculate and report statistical measures of performance: sensitivity, positive predictive value, and f-measure.

## References

1. Clancy, C.M. and J.R. Slutsky, *Commentary: a progress report on AHRQ's Effective Health Care Program*. Health Serv Res, 2007. 42(5): p. xi-xix.
2. *Institute of Medicine. Initial National Priorities for Comparative Effectiveness Research*. Washington DC: National Academics Press, 2009.
3. Congressional Budget Office. *Research on the Comparative Effectiveness of Medical Treatments: Issues and Options for an Expanded Federal Role*. Pub. No. 2975 Washington DC, 2007.
4. Smith, S., *Preface*. Medical Care, 2007. 45(10 Suppl 2): p. S1-S2.
5. Sox, H.C. and S. Greenfield, *Comparative effectiveness research: a report from the Institute of Medicine*. Ann Intern Med, 2009. 151(3): p. 203-5.
6. VanLare, J.M., P.H. Conway, and H.C. Sox, *Five next steps for a new national program for comparative-effectiveness research*. N Engl J Med, 2010. 362(11): p. 970-3.
7. Bloomrosen M, D.D., *Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association*. J Am Med Inform Assoc, 2008. 15(6): p. 715-722.
8. Centers for Disease Control and Prevention, *FOA: Enhancing Cancer Registry Data for Comparative Effectiveness Research* 2010, Atlanta, GA: CDC.
9. Sturmer, T., et al., *Nonexperimental Comparative Effectiveness Research Using Linked Healthcare Databases*. Epidemiology, 2011. 22(3): p. 298-301.