HSR Health Services Research

DOI: 10.1111/j.1475-6773.2010.01142.x
SPECIAL ISSUE: HEALTH SERVICES RESEARCH IN 2020

Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future

Cathy J. Bradley, Lynne Penberthy, Kelly J. Devers, and Debra J. Holden

Background. Research on pressing health services and policy issues requires access to complete, accurate, and timely patient and organizational data.
Aim. This paper describes how administrative and health records (including electronic medical records) can be linked for comparative effectiveness and health services research.
Materials and Methods. We categorize the major agents (i.e., who owns and controls data and who carries out the data linkage) into three areas: (1) individual investigators; (2) government sponsored linked data bases; and (3) public–private partnerships that facilitate linkage of data owned by private organizations. We describe challenges that may be encountered in the linkage process, and the benefits of combining secondary databases with primary qualitative and quantitative sources. We use cancer care research to illustrate our points.
Results. To fill the gaps in the existing data infrastructure, additional steps are required to foster collaboration among institutions, researchers, and public and private components of the health care sector. Without such effort, independent researchers, governmental agencies, and nonprofit organizations are likely to continue building upon a fragmented and costly system with limited access.
Discussion. Without the development and support for emerging information technologies across multiple health care settings, the potential for data collected for clinical and transactional purposes to benefit the research community and, ultimately, the patient population may go unrealized.
Conclusion. The current environment is characterized by budget and technical challenges, but investments in data infrastructure are arguably cost-effective given the need to reform our health care system and to monitor the impact of health reform initiatives.

Key Words. Administrative data, comparative effectiveness research, health information technology, electronic medical records, health services research, secondary data analysis

Research on pressing health services and policy issues requires access to complete, accurate, and timely patient and organizational data. However, in the United States, health-related datasets are created and held by diverse—often unrelated—public and private organizations and individual researchers. To overcome incomplete data from a single source, skilled researchers take months or years to acquire, link, and extract meaningful information from a myriad of secondary datasets. Through data linkage it is possible to get more complete information without the time and cost burden of additional and often duplicate primary data collection. Typical core datasets that are often linked and the information they contain are described in Table 1.

We provide an overview of commonly linked files and describe a generic process for linking datasets. We categorize the major agents (i.e., who owns and controls data and who carries out the data linkage) into three areas: (1) individual investigators, (2) government-sponsored linked databases, and (3) public–private partnerships that facilitate use and linkage of data owned and controlled by private organizations. These different agents shape whether and how readily disparate datasets can be accessed, linked, and mined by researchers. We also describe challenges that may be encountered in the linkage process, and the benefits of combining secondary databases with primary qualitative and quantitative sources.

Throughout the paper, we use cancer care research to illustrate our points. Cancer provides an excellent example because of its high prevalence and societal burden. As a result, several publicly and privately sponsored efforts have collaborated to create a data infrastructure that extends beyond traditional data systems. We conclude with recommendations to strengthen the existing data infrastructure. New strategies are needed to develop more accessible and comprehensive data systems to support the next generation of health services and policy research, including comparative effectiveness research.

Procedure for Linking Files

We describe five basic steps for linking databases: (1) identify the data sources that can be linked to answer a specific research question; (2) obtain the

Address correspondence to Cathy J. Bradley, Ph.D., Department of Healthcare Policy and Research, School of Medicine, Cancer Prevention and Control, Massey Cancer Center, Virginia Commonwealth University, Richmond, VA 23298-0203; e-mail: cjbradley@vcu.edu. Lynne Penberthy, M.D., is with the Department of Health Quality, School of Medicine, Virginia Commonwealth University, Richmond, VA. Kelly J. Devers, Ph.D., is with the Health Policy Center, Urban Institute, Washington, DC. Debra J. Holden, Ph.D., is with the Community Health Promotion Research, RTI International, Research Triangle Park, NC.

Table 1: Commonly Linked Databases

Category	*Description*	*Who Owns*	*Examples*
Claim files	(1) Organized at the claim level (2) Can be from a single payer (e.g., Medicare) or a single health system with multiple payers (3) Comprehensiveness varies. For example, if using Medicare files, data from third-party payers are not available or if using hospital discharge data, outpatient claims are not available	Insurers (federal, state insurance plans, private plans), providers (e.g., hospitals, medical groups) that submit claims	Medicare, Medicaid, private payers (e.g., Anthem), HCUP, and other hospital discharge data, providers
Disease registries	(1) Incidence based (2) Data on disease (site and stage of cancer), date of diagnosis, first course of treatment (surgery, radiation) (3) Includes patient-level data such as age and race	States, federal government, providers (chronic disease registries)	State cancer registries, Surveillance, Epidemiology, and End Results, birth defects registries, infectious disease registries
Survey	(1) Contains qualitative data—open-ended survey items (2) Self-reported patient- and provider-level information	Federal government, individual investigators	Health Retirement Survey, National Health Interview Survey
Provider files	(1) Provider or organization level (2) Provider characteristics-(e.g., ownership, size, staffing)	Government, organization professional associations	American Medical Association Physician Masterfile, American Hospital Association
Electronic medical records	(1) Designed for clinical care (2) Capacity to provide clinical information on each patient (3) Limited implementation and interoperability	Provider	Cerner, Epic, Eclypsis, Siemans, GE Centricity
Area level	(1) Organized at the county, ZIP code, census tract, or block	Government	Area resource file, U.S. Census data

continued

Table 1. *Continued*

Category	*Description*	*Who Owns*	*Examples*
	(2) Provides resource information (e.g., number of physicians, specialists, hospitals per 100,000 residents) and characteristics of geographic unit (e.g., median household income, racial composition, employment rates)		

necessary approvals, including institutional ethics boards, regulatory authorities, and funding sources; (3) select the variables that will be used to link the databases and individually clean the datasets; (4) determine the best method for linking databases and develop algorithms accordingly; and (5) evaluate the quality of the link between data sources.

Careful consideration of the research question, available data, and the strengths and limitations associated with data cannot be overemphasized. Data are expensive and time consuming to obtain, and they carry a high degree of responsibility in terms of protection, storage, and use. For the majority of research questions, the ideal dataset does not exist. Therefore, convenience and availability of secondary data should be weighed against a number of limitations, including relevance of the population covered and the ability to extract or impute the information needed.

Data linkage requires expertise in several areas, including knowledge of the datasets to be linked—their limitations and idiosyncrasies, skills in the use of linkage programs, and skills in statistical analysis and interpretation that comes from a multidisciplinary team. Database managers, programmers, and statisticians work collaboratively with health services researchers to resolve technical problems while keeping site of the research question.

In addition to the technical procedures and challenges involved in linking data, it is important to understand who owns the data and who will ultimately perform the linkage because these factors have a profound impact on whether the data can be accessed, in what time frame, at what cost, and whether other issues arise (e.g., legal constraints). Often researchers are based in universities or not-for-profit consulting firms and must seek funding to gain access to datasets owned by other organizations. This may require collaboration with someone employed by the organization. In addition, the acquisition of many commonly used databases (e.g., databases from Center for Medicare and Medicaid Services [CMS], National Center for Health Statistics Research Data Center) requires substantial fees.

Concerns about privacy led to policies that prevent records from being easily linked (Fellegi 1999). Therefore, a strong case for using the data and a detailed description of how it will be protected is required when obtaining Institutional Review Board (IRB) and other regulatory body (e.g., Privacy Boards) approvals (Safran et al. 2007). Different types of organizations (e.g., federal and state governments, private health plans, and providers) have varying interests in research and may be bound by different laws.

Once the appropriate regulatory approvals are obtained, the mechanics of data linkage can begin. At least one common identifier must exist between two datasets in order to link them. Common unique identifiers include Social Security Numbers (SSNs), Health Insurance Claim number (HICs), and Medical Record Numbers. These identifiers are used to link records at the patient level. Analogous identifiers at the hospital or area level are also available. Because of inevitable miscoding, the linkage can be improved by matching on variables such as sex, date of birth, names, addresses, and ZIP code in conjunction with a unique identifier. Race and ethnicity, even when available, are not good candidates for linkage because they are inconsistently and incorrectly reported across data sources. Once the linkage variables are selected, it is essential to ensure that these variables are as complete as possible and that no duplicate records exist in each dataset.

The next step is to select the best method or combination of methods to link datasets. The two methods used are deterministic and probabilistic matching. In deterministic matching, the investigator devises a series of steps that will be executed in a particular order to link two datasets. For example, the first step may be to attempt a complete match on SSN (or other unique identifier), sex, and month, day, and year of birth. The second step might be to match on less restrictive criteria, for example, the last four digits of the SSN, sex, and month, day, and year of birth. These steps are continued until as many records as possible are correctly linked between the two datasets.

A probabilistic matching process also uses identifiers to assess the likelihood that records from two datasets belong to a single identifier. Fellegi and Sunter (1969) formalized mathematical methods for considering a record "linked." Their seminal work defined a clear linkage rule that assigns a probability that two records from separate files represent the same person (or entity). Methods have since been developed that improve the accuracy and efficiency of Fellegi and Sunter's original work (see e.g., Winkler 1993 and Jaro 1995). Probabilistic linkage requires investment in software that will perform the match.

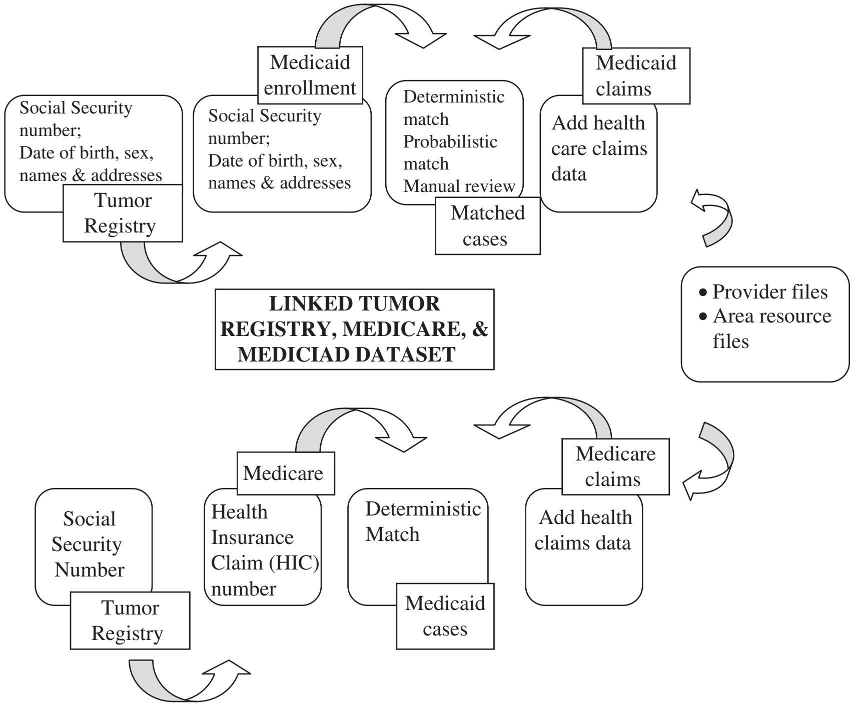
The final step is to evaluate the quality of the match between two datasets. During the evaluation process, records may be manually reviewed to determine whether the matching algorithm performed correctly and whether lower quality matches are valid. It may be useful to write programs to evaluate the quality of less than perfectly matched records. These algorithms reduce the time spent manually reviewing records and improve the quality of acceptable matches. This process can be lengthy and should be considered in the timeline for the project.

The following sections describe examples from individual investigator, government, and public–private partnerships that have linked data for the purposes of health services research. The challenges and limitations are described within the technical framework outlined in the preceding section.

INVESTIGATOR-INITIATED LINKED DATABASES

Bradley et al. (2007) studied the impact of dual Medicaid and Medicare eligibility on cancer treatment and outcomes. Four core datasets were used: Medicare and Medicaid claims files and a state Tumor Registry data, which were later supplemented by provider, and area-level data files. To obtain these data, the investigators worked closely with state health department officials in vital statistics and Medicaid. Approval was required by the director of the state department of community health, and IRB approval was required at the state and university. It took 2 years to acquire all approvals and the data. These files were ultimately linked by the state health department employees (primarily staff in the cancer registrar's office) because researchers were not allowed access to certain identifying patient information. The methods (in this case, both deterministic and probabilistic) used to link the files were jointly negotiated between the researchers and state employees. The project may not have been possible in other states, which vary in how they process and maintain Medicaid claim files and in their interest in and resources for working with researchers. On a practical level, many states use identification numbers that do not neatly correspond with HICs or SSNs and use outside vendors to process their Medicaid claim files (Prela et al. 2009).

Figure 1: Example of How Data Are Linked to Create a Matched Tumor Registry, Medicaid, and Medicare Dataset with Provider- and Area-Level Files


Medicaid enrollment
Medicaid claims
Social Security number; Date of birth, sex, names & addresses
Tumor Registry
Social Security number; Date of birth, sex, names & addresses
Deterministic match Probabilistic match Manual review
Matched cases
Add health care claims data
LINKED TUMOR REGISTRY, MEDICARE, & MEDICIAD DATASET
• Provider files
• Area resource files
Medicare
Medicare claims
Social Security Number
Tumor Registry
Health Insurance Claim (HIC) number
Deterministic Match
Medicaid cases
Add health claims data

For this project, the State Director of the Office of Vital Statistics sent a file of SSNs from the Tumor Registry to the CMS to link with HICs. CMS used the beneficiary's SSN to link subjects to a particular HIC. CMS performed a deterministic link and returned an SSN to HIC conversion file along with a cross-reference file for beneficiaries with multiple HICs. Separate claims files were then obtained for inpatient, outpatient, physician and supplier, skilled nursing facility, durable medical equipment, and hospice services provided to beneficiaries enrolled in Medicare fee-for-service. The cost of these data approached U.S.$100,000 and required additional investment in computer storage and equipment compliant with the Health Insurance Portability and Accountability Act. Once the linked data and claims were assembled, provider- and area-level files were added. Figure 1 illustrates how the files were linked.

Together, these data represent the type of linked data systems an investigator would create independently, often using federal grant dollars.

The advantages of using claims data from social insurance programs and then linking them to other data sources include the vast number of beneficiaries, the diverse geographic areas and providers represented, and the relative accessibility of the data, possibly allowing for replication and reproducibility of studies. Claims data can be synergistic with other data as in the case of claims and registry data. These advantages are offset by several key limitations, including the exclusion of patients enrolled in private health plans, being prohibitively expensive to obtain, and the complexity required to assemble and validate linked data. Finally, once research is completed by the individual investigator (or others initially authorized to use the dataset), these expensive linked systems must be destroyed (as a requirement of Data Use Agreements) and are not available for other research. As a result of these limitations, the return on investment in investigator-initiated linked data may be limited. The next section discusses government-sponsored linked datasets, which overcome some of the limitations of investigator-initiated linked data.

GOVERNMENT-SPONSORED LINKED DATASETS

Government-sponsored linked datasets may offer a methodologically sound and cost-effective alternative to investigator-linked data. Federal and state governments have more datasets available to them along with extensive identifying information. Federally sponsored linked projects can use a standard methodology for linking and assembling data and can make deidentified data available to researchers. A prominent example in cancer research is the National Cancer Institute's (NCI's) linked Surveillance, Epidemiology, and End Results (SEER)-Medicare files.

The SEER program is funded by NCI to collect cancer incidence, first course of surgery or radiation therapy treatment, and survival data from population-based cancer registries covering approximately 26 percent of the U.S. population. The SEER program operates registries in 10 states and the metropolitan areas of Detroit and Atlanta, 10 rural counties in Georgia, and the 13-county Seattle-Puget Sound area. NCI and CMS update the linked SEER-Medicare files every 2 years using an algorithm that includes subject SSN, name, sex, and date of birth (Warren et al. 2002). NCI retains ownership of the data and releases it for approved research studies that ensure the confidentiality of the patients and providers in the SEER areas. Once approval is obtained, the investigator purchases the dataset. The combination of SEER and Medicare claims overcomes many of the limitations of each separate

Table 2: Examples of Government-Sponsored Linked Datasets

Dataset	*Purpose*	*Population Covered*	*Has Been Linked to . . .*	*Restrictions*	*Website*
Health and Retirement Study	Longitudinal study (every 2 years) income, health, and employment dynamics	National sample of household population age 51 and older and their spouses	Social Security earnings data; Medicare claim files	Restrictions apply to Social Security and Medicare claim files	http://hrsonline.isr.umich.edu/index.php
Health Care Cost and Utilization Project	Longitudinal study all-payer information from hospital discharge abstract (patient diagnoses, treatment, charges, and patient demographics). Data on emergency departments and ambulatory surgery centers	National sample database and statewide databases also available for patients in all age groups	Organization files (e.g., AHA hospital characteristics and/or Medicare cost reports); Area Resource File	Agency for Health care Research and Quality facilitates linkage with AHA and Medicare cost reports. Researchers need to link with ARF or other data sources	http://www.hcup-us.ahrq.gov/
Medicare Current Beneficiary Survey	Provides information on health status, health care use and expenditures, health insurance coverage, and socioeconomic characteristics for Medicare beneficiaries	National representative sample of aged, disabled, and institutionalized Medicare beneficiaries	Medicare claim files		http://www.cms.hhs.gov/MCBS/

Medical Expenditure Panel Survey	Panel data on use of health care services, frequency of use, cost, and payment type. Also includes information on cost, scope, and breadth of health insurance held by and available to U.S. workers	National representative subsample of households; sample of private and public sector employers on the health insurance plans they offer their employees	National Health Interview Survey	Insurance component data files are not available for public release	http://www.meps.ahrq.gov/mepsweb/
Minimum Dataset	Nursing home-level data for clinical assessment of all residents in Medicare- or Medicaid-certified nursing facilities and provides information to providers to compare their quality of care to state standards	All residents of Medicare- or Medicaid- certified nursing facilities	Medicare claim files	Individual facility information not available to public	http://www.cms.hhs.gov/MinimumDataSets20/
National Health Interview Survey	Cross-sectional household interview survey to monitor the health of the U.S. population through information on health characteristics	Civilian noninstitutionalized population residing in the United States at the time of the interview	Mortality data, Medicare Enrollment and Claims data, Social Security Benefit History data; MEPS		http://www.cdc.gov/nchs/nhis.htm http://www.cdc.gov/nchs/data_access/data_linkage_activities.htm

continued

Table 2. *Continued*

Dataset	*Purpose*	*Population Covered*	*Has Been Linked to . . .*	*Restrictions*	*Website*
Surveillance, Epidemiology, and End Results (SEER) Registry-Medicare	Registry of all incident cancer cases from 10 states and the metropolitan areas of Detroit and Atlanta, 10 rural counties in Georgia, and the 13-county Seattle-Puget Sound area	Approximately 26% of the U.S. population (SEER) linked with Medicare fee-for-service claims files of patients age 65 and older	Medicare claims files; limited American Medical Association, AHA, and census data available	All requests must be approved by National Cancer Institute; additional data cannot be linked by the investigator	http://seer.cancer.gov

AHA, American Hospital Association.

dataset and also links with data found in the ARF, Census, American Hospital Association, and American Medical Association datasets. Because these data are assembled for research purposes, they have been used widely and have contributed much to what is known about cancer patterns of care and outcomes. Table 2 lists examples of other government-sponsored linked datasets, along with their access information, and summarizes advantages and disadvantages of each dataset.

Public Private Partnerships in Creating and Using Data Systems

A significant amount of health-related data are owned and controlled by private organizations, which provides an opportunity for public and private organizations (e.g., health plans, organized delivery systems, professional associations) to collaborate and partner with researchers in universities or other settings (e.g., not-for-profit consulting firms). These partnerships extend to populations that are not covered by government-sponsored health plans and include the collection, linkage, and use of a wide range of quantitative or qualitative data. Without these partnerships, timely and comprehensive health information would not be available from private organizations and their patient populations to answer pressing health services and policy research questions, including the degree to which specific programs or initiatives to advance health care are working. This information is vital to the future and success of comparative effectiveness research.

An example of a public–private partnership to stimulate improvements in cancer research and care is the NCI Community Cancer Centers Program (NCCCP). The NCCCP is a 3-year pilot program to test the concept of a network of hospital-based community cancer centers (CCCs) located in 14 states with the goals of bringing more Americans into high-quality multidisciplinary cancer care; increasing participation in clinical trials; exploring standards for collecting and storing cancer research specimens; reducing cancer health care disparities; and improving information sharing among CCCs, including greater use of electronic medical records (EMRs) and other health information technology (Fennell 2008; Clauser 2009; Clauser et al. 2009; Johnson et al. 2009; NCI 2009a, b). The NCI funded a total of 10 programs: eight individual hospital-based CCCs and CCCs in two hospital systems that include 16 hospitals and their cancer centers. The NCI, in conjunction with its private partners (e.g., participating CCCs, professional associations like the

American College of Surgeons [ACOS] Commission on Cancer and an independent research evaluation team) are collecting, linking when appropriate and possible, and analyzing a variety of primary and secondary quantitative and qualitative data to manage the NCCCP (Holden et al. 2009).

Given the objectives and nature of the NCCCP, the primary units of analysis are the 16 participating CCCs and the pilot program overall. Therefore, quantitative (primary and secondary) and qualitative data at the patient, provider, and area levels are collected and linked to assess which CCCs have made the most progress, why, and the extent to which the NCCCP has achieved its aims and at what cost. Such information is critical for identifying the organizational requirements and environmental conditions necessary to effectively implement the program and sustain it over time. Primary quantitative, patient-level data consist of a patient survey, whose sampling frame was developed from the CCCs' tumor registry, and detailed clinical information collected within 30 days of a patient visit or diagnosis on quality measures specific to breast and colorectal cancer care (i.e., via the ACOS Rapid Quality Reporting System). The NCI has addressed confidentiality and privacy issues with NCCCP data. Primary, quantitative provider-level data consist of a comprehensive CCC survey (baseline, interim, and final) and detailed cost data to examine the start up and incremental program-related costs to the CCC. Secondary, quantitative provider-level data consist of reports the CCCs submit to ACOS for accreditation. Lastly, the ARF is also linked to understand key aspects of the environment in which the CCCs operate (e.g., income/education level and race/ethnicity of local population, percentage uninsured, managed care penetration). These quantitative data at multiple levels are complemented by a rich array of qualitative data, including data from patient focus groups, in-person open-ended interviews with hospital and CCC leadership staff, and CCC and NCCCP pilot documentation.

Collectively, these quantitative and qualitative data represent many desired components of a linked dataset. They are owned and controlled by NCI but will be used by three different groups: NCI staff responsible for managing the program; participating CCCs to benchmark their performance and learn best practices from other sites; and the evaluation team. In addition, all NCCCP sites are working to expand their linkages with the NCI-designated cancer centers and their associated investigators, resulting in additional research opportunities.

If NCCCP continues beyond the pilot phase, a subset of these data could be collected on an ongoing basis to support management of the NCCCP and provide useful information for program improvement and future research on the impact of finance, organization, and delivery changes on cancer care and research. NCCCP pilot accomplishments to date in EMRs and Health

Information Technology strengthen the cancer data infrastructure and support related future quality improvement and possibly comparative effectiveness research initiatives (NCI 2008a, 2009b).

Expansion Areas for Linked Data Systems

This section describes opportunities for future data linkages. EMRs, for example, have been a focus of attention for their potential use in research. The Institute of Medicine (United States), Committee on Data Standards for Patient Safety, Institute of Medicine (United States), and Board on Health Care Services (2003) developed a list of potential functionalities, which might be used to determine the value or utility of a particular EMR[1] for research purposes. For example, EMR should have a well-defined and extractable set of patient demographics across the entire health care system; a "clinical dashboard" for use in reporting key quality indicators or legally mandated conditions (e.g., cancer, infectious diseases); and a structured template-based system for capturing information in specific clinical areas that may enhance the utility of the data for research purposes (McLeod 2007).

As of this writing, EMR use and interoperability is not sufficiently developed to permit routine use for research purposes. The penetration of EMR use in acute care hospitals is low, with only 1.5 percent of hospitals having a comprehensive EMR (Jha et al. 2009). Nevertheless, more than 75 percent of hospitals have electronic laboratory and radiologic reporting systems. This suggests that while the EMR does not currently provide a likely source of research data, targeted components of the EMR may have a higher value for extraction and linkage with other data sources for research purposes. For example, leveraging the high penetration of electronic laboratory records such as tumor marker test results could be used to rapidly identify cancer patients with recurrence. These data might be linked with chemotherapy treatment information obtained from billing data and also to cancer registry data to monitor patient's diagnosis and response to therapy.

Linkages with components of an EMR system are only likely to occur within or across a health care entity or group of related hospitals sharing the same fully interoperable EMR systems. If a researcher were to link these same data from multiple health care entities, the linkage process becomes more complex. The linkage could be performed using SSN; however, extracting the data in a similar and complimentary format is challenging and may require manual data abstraction and entry based on the text information in the source documents. Studies have demonstrated some success in using natural language processing (NLP) systems to identify disease. As they become more reliable and widely available, the use of NLPs is likely to greatly enhance the value of text-based messages (McCowan, Moore, and Fry 2006; McCowan et al. 2007; Pakhomov et al. 2007; Savova et al. 2008).

For many diseases, diagnosis and treatment are provided exclusively in the outpatient setting. Thus, leveraging interoperability of existing systems such as physician billing data also offers an exciting area for expanding the existing data infrastructure necessary for health services and comparative effectiveness research. Access to physician billing data permits the capture of relevant information on treatment in a standardized format. Billing data have high sensitivity and specificity and are valid for specific types of treatment such as chemotherapy. Physician practice data cover the treated population regardless of payer. Such a system might be cost-effective through the use of automated screening of electronically submitted claims. These data would provide information on treatment in patients across insurers that would be otherwise missed through linkages to claims data from specific payers (e.g., Blue Cross Blue Shield). Working with large specialty practices such as hematology/oncology would provide supplemental information that could enhance the utility of tumor registry data by providing more complete and longitudinal data on patients for a geographic region. Larger groups such as network practices have the potential to provide data on a scale that would permit analysis of population subgroups within a region.

Likewise, pharmacy data could significantly contribute to our knowledge of outpatient therapies, particularly in emerging treatment categories in which there is limited comparative effectiveness research. All pharmacies in 38 states (with an additional 11 in process) report electronically to a central data repository under a mandate to report controlled substances. This system could be expanded to include other prescription information. States mandate that physicians and other health care providers report cancer treatments; expanding the mandate to pharmacies might be a logical step that could supply information that can be used with other data to address important policy issues. A critical strength of this system is that pharmacies report both insured and self-pay prescriptions.

Recommendations

This section suggests high-priority recommendations to enhance data systems and their accessibility to researchers. Most recommendations focus on a systematic and centralized approach to maximizing existing data systems and how to move forward so that important gaps in the data infrastructure are filled.

Develop a Plan

Convene a panel of government and nongovernment experts to develop a comprehensive plan for expanding the warehouse of linked files. This panel should be sponsored by the National Institutes of Health (NIH) and Agency for Health Care Research and Quality (AHRQ). The panel would be tasked with, but not limited to, the following activities:

- Inventorying all publicly sponsored health-related datasets. These datasets should include federal and state sponsored projects.
- Examining how these datasets could be linked and identify barriers that prohibit cooperation, agreement, and data sharing.
- Identifying areas of overlap and duplication in data collection.
- Developing a set of priorities and recommendations for data sharing and linkage.
- Identifying an appropriate custodian for gathering, linking, deidentifying, storing, and distributing linked data.

Remove Barriers

Convene a consortium of interagency governmental and private health care organizations to identify policies and practices (e.g., mandates to destroy linked datasets, prohibitions on states linking CMS data with cancer registry data) that constrain the use of data. The consortium will be tasked with developing policies to remove barriers to linking and making data available for research. The use of data from genomics research and others (e.g., cancer Biomedical Informatics Grid) is moving toward open policies and availability. Policies developed in these domains may have relevance to the broader field of health services research.

Develop Standards

Convene a scientific, technically oriented task force consisting of government and private sector members to develop a robust set of policies, standards, and best practices for linking and using secondary data so that issues around quality, confidentiality, data storage, and use can be resolved. Specific recommendations regarding methods used to match records across files and acceptable match rates should emerge from the scientific committee. The task force should be convened by the NIH and AHRQ.

Capitalize on State Systems

Support data collection and linkages at the state level. States mandate the collection of many types of health-related data, including disease registries, all-payer databases, Medicaid claims, and treatment data. With support, individual states could develop a centralized, linked warehouse for their data. These centralized and consolidated data warehouses would benefit researchers and state agencies as well. Once several states develop high-quality warehouses, these data can be merged across states for regional- and national-level analyses. States with the greatest degree of experience and resources could compete on an award to support the development of a health data warehouse and take the lead to establish a structure for other states to emulate. Support from the federal government is needed to guide a demonstration-like project and to ensure that states coordinate their efforts.

Collect and Integrate Qualitative Data

Support the collection of primary quantitative data (e.g., patient survey) and qualitative data that complement secondary sources. Qualitative data identify trends that are not detected in fixed data sources and can improve understanding of important aspects of the environmental, social, and organizational context that affects policy or program implementation (Rundall, Devers, and Sofaer 1999). As in the NCCCP example, qualitative data provide unique, complementary insights such as promising strategies, practices, and tools that help accelerate and sustain progress toward NCCCP goals and the impact of the NCCCP on providers and patients.

During the interim in which the centralized task forces and scientific committees are working, the federal government could offer more hands-on user training to facilitate broader use of existing datasets. In addition, clearinghouses and/or distributors to assist researchers and facilitate access to government-sponsored databases are needed. The Research Data Assistance Center, which provides free assistance to researchers interested in using Medicare and Medicaid data, is an excellent model for how data clearinghouses would function.

To fill the gaps in the existing data infrastructure, additional steps are required to foster collaboration among institutions, researchers, and public and private components of the health care sector. Without such effort, independent researchers, governmental agencies, and nonprofit organizations are likely to continue building upon a fragmented and costly system with limited access. Without the development and support for emerging information technologies across multiple health care settings, the potential for data collected for clinical and transactional purposes to benefit the research community and, ultimately, the patient population may go unrealized. Much of the population's health-related data resides in the private domain. A concerted effort is required to integrate private data sources with public sources.

CONCLUSIONS

Researchers have turned to linked data systems to enhance existing data, reduce the cost of data acquisition, and to avoid duplicate primary data collection. We provide an overview of how to link these data and describe other data systems that have been linked. We also illustrate how existing data sources could be extended by thoughtful supplements and integration and how public and private partnerships hold potential for creating a complementary and comprehensive data infrastructure accessible to health services and policy researchers nationwide. The current environment is characterized by budget and technical challenges, but investments in data infrastructure are arguably cost-effective given the need to reform our health care system and to monitor the impact of reform initiatives.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This paper was commissioned by AcademyHealth with support from The Commonwealth Fund and The Robert Wood Johnson Foundation.

Disclosures: None.
Disclaimers: None.

NOTE

1. Personal health records (PHRs) and Regional Health Information Organizations (RHIOs) hold promise for expanding and enhancing data for health services and policy research. For an overview of PHRs and related data linkage and privacy issues, see: Halamka, Mandl, and Tang (2008), Tang et al. (2006), Tang and Lansky

(2005). For an overview of RHIOs and related data linkage and privacy issues, see: Hollar (2009), Miller and Miller (2007), Frohlich et al. (2007), Holmquest (2007), and Glaser (2007).

REFERENCES

Bradley, C. J., C. W. Given, Z. Luo, C. Roberts, G. Copeland, and B. A. Virnig. 2007. "Medicaid, Medicare, and the Michigan Tumor Registry: A Linkage Strategy." *Medical Decision Making* 27 (4): 352–63.

Clauser, S. B. 2009. "National Cancer Institute Partnerships in Quality-of-Care Research." *Cancer Control.* 16 (4): 283–92.

Clauser, S. B., M. R. Johnson, D. M. O'Brien, J. M. Beveridge, M. L. Fennell, and A. D. Kaluzny. 2009. "Improving Clinical Research and Cancer Care Delivery in Community Settings: Evaluating the NCI Community Cancer Centers Program." *Implementation Science* 4: 63.

Fellegi, I. 1999. "Record Linkage and Public Policy—A Dynamic Evolution." In *Record Linkage Techniques 1997*, edited by W. Alney, and B. Jamerson, pp. 3–12. Washington, DC: National Academy Press.

Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association.* 64 (328): 1183–210.

Fennell, M. L. 2008. "The New Medical Technologies and the Organization of Medical Science and Treatment. Editorial." *Health Services Research* 43 (1, part I): 1–9.

Frohlich, J., S. Karp, M. D. Smith, and W. Sujansky. 2007. "Retrospective: Lessons Learned from the Santa Barbara Project and Their Implications for Health Information Exchange." *Health Affairs* 26 (5): w589–91.

Glaser, J. 2007. "The Advent of RHIG 2.0: The Country's Strategy of Creating Clinical Data Exchanges Is about to Undergo a Difficult Shift from RHIO 1.0 to RHIO 2.0." *Journal of Healthcare Information Management* 21 (3): 7–9.

Halamka, J. D., K. D. Mandl, and P. C. Tang. 2008. "Early Experiences with Personal Health Records." *Journal of the American Medical Informatics Association* 15 (1): 1–7.

Holden, D. J., K. J. Devers, L. McCormack, K. Dalton, S. Green, and K. Trieman. 2009. *Evaluation Design Report for the National Cancer Institute's Community Cancer Centers Program.* Final Report. Research Triangle Park, NC: National Cancer Institute, RTI International.

Hollar, D. W. 2009. "Progress along Developmental Tracks for Electronic Health Records Implementation in the United States." *Health Research Policy and Systems* 7: 3.

Holmquest, D. L. 2007. "Another Lesson from Santa Barbara." *Health Affairs* 26 (5): w592–4.

Institute of Medicine (United States), Committee on Data Standards for Patient Safety, Institute of Medicine (United States), and Board on Health Care Services. 2003. *Key Capabilities of an Electronic Health Record System: Letter Report.* Washington, DC: The National Academies Press.

Jaro, M. 1995. "Probabilistic Linkage of Large Public Health Data Files." *Statistics in Medicine*. 14: 491–8.

Jha, A. K., C. M. DesRoches, E. G. Campbell, K. Donelan, S. R. Rao, T. G. Ferris, A. Shields, S. Rosenbaum, and D. Blumenthal. 2009. "Use of Electronic Health Records in U.S. Hospitals." *New England Journal of Medicine* 360 (16): 1628–38.

Johnson, M. R., S. B. Clauser, J. M. Beveridge, and D. M. O'Brien. 2009. "Translating Scientific Advances into the Community Setting: The National Cancer Institute Community Cancer Centers Program Pilot." *Oncology Issues* 24: 24–28.

McCowan, I., D. Moore, and M. J. Fry. 2006. "Classification of Cancer Stage from Free-Text Histology Reports." *Conference Proceedings—IEEE Engineering in Medicine and Biology Society* 1: 5153–6.

McCowan, I. A., D. C. Moore, A. N. Nguyen, R. V. Bowman, B. E. Clarke, E. E. Duhig, and M. J. Fry. 2007. "Collection of Cancer Stage Data by Classifying Free-Text Medical Reports." *Journal of the American Medical Informatics Association* 14 (6): 736–45.

McLeod, R. S. 2007. "Comparison of Data Extraction from Standardized versus Traditional Narrative Operative Reports for Database Related Research and Quality Control." *Surgery* 142 (3): 420–1.

Miller, R. H., and B. S. Miller. 2007. "The Santa Barbara County Care Data Exchange: What Happened?" *Health Affairs* 26 (5): w568–80.

National Cancer Institute (NCI). 2008a *caBIG Fact Sheet.* NIH Publication No. 08-6457 [accessed on June 23, 2009]. Available at https://cabig.nci.nih.gov/overview/caBIG_Fact_Sheet.pdf

National Cancer Institute (NCI).f 2008b. *NCI Community Cancer Centers Program Pilot: 2007–2010: The First Year.* NIH Publication No. 138 [accessed on June 23, 2009]. Available at http://ncccp.cancer.gov/About/Progress.htm

National Cancer Institute (NCI). "NCCCP Pilot Program Summary" [accessed on May 14, 2009a]. Available at http://www.cancer.gov/researchandfunding/ncccp-pilot-program

National Cancer Institute (NCI). "NCI Community Cancer Centers Program—About NCCCP Overview" [accessed on May 14, 2009b]. Available at http://ncccp.cancer.gov/About/index.htm

Pakhomov, S., S. A. Weston, S. J. Jacobsen, C. G. Chute, R. Meverden, and V. L. Roger. 2007. "Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure." *American Journal of Managed Care* 13 (6, part 1): 281–8.

Prela, C. M., G. A. Baumgardner, G. E. Reiber, L. V. McFarland, C. Maynard, N. Anderson, and M. Maciejewski. 2009. "Challenges in Merging Medicaid and Medicare Databases to Obtain Healthcare Costs for Dual-Eligible Beneficiaries: Using Diabetes as an Example." *Pharmacoeconomics* 27 (2): 167–77.

Rundall, T. G., K. J. Devers, and S. Sofaer. 1999. "Overview of the Special Supplement Issue." *Health Services Research* 34 (5, part 2): 1091–9.

Safran, C., M. Bloomrosen, W. E. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, D. E. Detmer, and P. Expert. 2007. "Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper." *Journal of the American Medical Informatics Association* 14 (1): 1–9.

Savova, G. K., P. V. Ogren, P. H. Duffy, J. D. Buntrock, and C. G. Chute. 2008. "Mayo Clinic NLP System for Patient Smoking Status Identification." *Journal of the American Medical Informatics Association* 15 (1): 25–8.

Tang, P. C., J. S. Ash, D. W. Bates, J. M. Overhage, and D. Z. Sands. 2006. "Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption." *Journal of the American Medical Informatics Association* 13 (2): 121–6.

Tang, P. C., and D. Lansky. 2005. "The Missing Link: Bridging the Patient–Provider Health Information Gap." *Health Affairs* 24 (5): 1290–5.

Warren, J. L., C. N. Klabunde, D. Schrag, P. B. Bach, and G. F. Riley. 2002. "Overview of the SEER-Medicare Data: Content, Research Applications, and Generalizability to the United States Elderly Population." *Medical Care* 40 (8, suppl): 3–18.

Winkler, W. 1993. "Improved Decision Rules in the Fellegi–Sunter Model of Record Linkage." *American Statistical Association 1993 Proceedings of the Section on Research Methods*, pp. 273–8.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.