March 22, 2010

Connecting the Dots: The Promise of Integrated Data Systems for Policy Analysis and Systems Reform

Dennis P Culhane, *University of Pennsylvania*
John Fantuzzo, *University of Pennsylvania*
Heather L Rouse, *University of Pennsylvania*
Vicky Tam, *University of Pennsylvania*
Jonathan Lukens, *University of Pennsylvania*

INTELLIGENCE FOR SOCIAL POLICY

Transforming Education, Health and Human Services through Integrated Data Systems

March 22, 2010
Vol. 1, No. 3

University of Pennsylvania T: (215) 573-7823 F: 215 (215) 573-2099 www.isppenn.org

Connecting the Dots: The Promise of Integrated Data Systems for Policy Analysis and Systems Reform

Dennis P. Culhane, John Fantuzzo, Heather L. Rouse, Vicky Tam & Jonathan Lukens

Abstract

This article explores the use of integrated administrative data systems in support of policy reform through interagency collaboration and research. The legal, ethical, scientific and economic challenges of interagency data sharing are examined. A survey of eight integrated data systems, including states, local governments and university-based efforts, explores how the developers have addressed these challenges. Some exemplary uses of the systems are provided to illustrate the range, usefulness and import of these systems for policy and program reform. Recommendations are offered for the broader adoption of these systems and for their expanded use by various stakeholders.

Introduction

Optimizing the coordination among integrated data systems can be a matter of life or death. When Donald F. Kettl, University of Maryland Dean of the School of Public Policy and former Director of University of Pennsylvania's Fels Institute of Government, looked at the response of national, state and local governments to Hurricane Katrina, he concluded that the inadequate response to that massive crisis was not due to the failure of any one system—but rather the result of "problems of coordination (of information) at the interface between multiple systems." (*Is the Worst Yet to Come?* Annals of the American Academy of Political and Social Science, vol. 604, March 2006.)

This country faces a multitude of problems as complex as the response to Hurricane Katrina, if not as dramatic. Delivery systems in such diverse areas as health, education and criminal justice often do not or cannot share information in a way that could improve services, both for individuals and on a larger scale. Building capacities for timely, data-based decision-making across multiple systems will not only result in greater efficiencies in service delivery; it will also benefit policymakers, who can use such integrated data to answer critical policy and program questions: what works, for whom, and at what cost. The integration of administrative data across service agencies has been identified as the next frontier for generating quality evidence to inform public policy and system reform. (Duran, Wilson, & Carroll, 2005; Hotz, Goerge, Balzekas, & Margolin, 1998)

The complex problems facing citizens in the US require a thoughtful consideration of how we can build capacities for data-based decision-making across diverse service delivery systems. Policy makers need timely data integrated across multiple systems in order to coordinate the services that are needed by citizens, including many vulnerable populations. These integrated data are needed to describe the conditions of program participants and the services they receive. They are also needed to answer the critical policy and program questions of what works, for whom and at what cost.

As a result of these pressing needs, administrative databases provide a powerful source of information for research and policy analysis. Because they track the front-line activities of public agencies, administrative data are directly relevant to program design, management, and evalua-

tion. Administrative records, routinely gathered and maintained, provide tremendous opportunities for longitudinal, population-based research, with real-time or nearly real-time data. Broadly, a program's administrative database can be used to identify

- the prevalence and patterns of service utilization within a given agency,
- the risk and protective factors associated with program use, and
- the costs associated with various patterns of utilization.

But people who use public programs are often users of other programs, and are at different developmental points in the course of their lives. Public agencies have much to gain by understanding how their collective activities could be leveraged to maximize outcomes and to optimize the use of resources, both across programs and over time.

Thus, the integration of administrative data systems provides potentially even more compelling information on patterns of multi-system program use, costs and outcomes. Here are a few ways that such data can be used.

- Interventions or program investments in one domain (e.g., housing stabilization) can be designed and evaluated to reduce the use of costly or inappropriate services in another area (e.g., health care).
- Programs can be designed to target particular subpopulations of program users (e.g., preschool children) who are known to have identified antecedents of care in other systems (e.g., child welfare).
- Policy analysts can use these data to identify which programs in one area (e.g., afterschool programs) may have the most significant long-term gains as measured by program outcomes in other areas, and across the life-course (e.g., reduced teen births or transmission of STDs).

Perhaps as importantly as the results that it can provide, such research might be possible in months rather than years, and at a fraction of the cost as compared to longitudinal research based on primary data collection.

Encouraged by the prospect of such gains in program efficiency and in improved outcomes for program consumers, several organizations throughout the United States have independently developed their own integrated data systems (IDS). These are projects led by state governments, local governments, and universities. Without any national program structure or even published guidance, these systems have evolved within their own contexts to meet the information and research needs of their partners.

We looked at eight of these diverse exemplary systems to extend our understanding of the current state of the development and use of IDS. As a guiding framework for our inquiry, we distinguished four broad sets of challenges facing those who develop, implement and use such systems: legal, ethical, scientific, and economic. We surveyed these eight existing IDS, leading to a preliminary picture of the range of public agencies providing data to these efforts, and some of the distinctive uses to which these data have been put. From these findings, we offer recommendations for how both current and future data systems could be leveraged to answer some of the most important public policy questions facing our society today. And we consider how other communities and other public policy stakeholders can benefit from the experience of these innovators.

Background: Challenges of Integrated Data Systems

Legal Challenges

When integrated data systems are used for research, a number of complex legal issues must be considered relating to the privacy of persons within these systems. The rights to use various types of data for research are regulated by federal law, state law, and public access policies. At each level of government, there are provisions that permit access and integration across these administrative data systems.

Federal Law

The Privacy Act of 1974, 5 U.S.C. § 552a (2000), is the omnibus "code of fair information practices" that regulates the collection, maintenance, use, dissemination, and disposition of personal information by the federal government. The Privacy Act is designed to balance the government's need to maintain information about individuals with the rights of individuals to be protected against unwarranted disclosure of their personal information.

Two other legislative enactments specifically address federal legislative guidelines for the protection of

individual health records and education records—the Health Insurance Portability and Accountability Act of 1996 and the Family Educational Rights and Privacy Act of 1974. Other federal laws protect privacy of tax records, census data, child support enforcement, drivers licensing information, banking and financial records, etc., but because such kinds of data are typically not included in integrated data systems that support service coordination and planning, we will not cover them here.

HIPAA

Standards for protecting the privacy of individually identifiable health information were established by the United States Department of Health and Human Services (HHS), implementing regulations promulgated under the Health Insurance Portability and Accountability Act of 1996 (HIPAA). These regulations address the use and disclosure of protected health information by covered entities including health insurance plans, health care clearinghouses, and health care providers who transmit electronic claims subject to HIPAA's administrative simplification standards. Protected health information must be directly linked to identifying information about an individual (e.g., name, social security number) (45 C.F.R. §§ 160.102, 160.103). A major goal of HIPAA is to assure that individuals' health information is properly protected while allowing for the flow of health information to promote high quality health care and protect the public's health and wellbeing.

HIPAA protects private health information and creates provisions for the use of such information to improve public services and policy making. The law provides the authority for public health agencies to engage in partnership agreements with researchers, who serve as business agents on behalf of the agency. These partnership agreements (known as Business Associate agreements under the law) are contracts between researchers and service agencies for the completion of agency-designed research. They provide for the completion of internal research projects to support policy and planning by allowing agencies to contract with experts to complete the work.

A second set of provisions within the federal privacy legislation speaks to the disclosure of individual records to external researchers for the purposes of statistical inquiry (5 U.S.C. § 552a). These stipulations permit the sharing of records to a third party who has provided the agency with adequate written assurance, in advance, that the record will be used solely as statistical research, and that the record will be transferred in a form that is not individually identifiable. Such research is considered one of the allowable categories of "public interest and benefit activities," so long as the research is designed to develop or contribute to generalizable knowledge (45 C.F.R. § 164.501). Several provisions are also provided within HIPAA for the use of identified data for research by covered entities and their business associates. A final stipulation indicates that there are no restrictions on the use or disclosure of de-identified health information (45 C.F.R. §§ 164.502(d)(2), 164.514(a) and (b)). The de-identification process involves the removal of specified data elements pertaining to the individual, as well as the individual's relatives, household members, and employers.

FERPA

The Family Educational Rights and Privacy Act of 1974 (FERPA; 20 U.S.C. § 1232g) protects information contained in public education records about parents and students. Similar to the HIPAA regulations, FERPA states that public education agencies may not institute any policy permitting the release of personally identifiable records without prior written consent from parents, or from students who have reached the age of majority. As with HIPAA, there are explicit exceptions to the "prior written consent" rule. One of these exceptions is the provision for sharing of information with organizations conducting studies for or on behalf of the educational agency or institution. Such studies must serve an administrative purpose of the educational agency, including developing, validating, or administering predictive tests, administering student aid programs, and improving instruction. These studies must be conducted in a manner that does not permit the personal identification of students and their parents, and researchers must agree that the information will be destroyed when no longer needed for the purpose for which it is provided (20 U.S.C. § 1232g (b)(1)(D)).

State Law

In addition to the explicit federal regulations for the protection of health and education data through HIPAA and FERPA, states are required to protect the privacy of children and families served by other public service agencies, such as child welfare, housing and homelessness, and juvenile justice. In these areas, state and local governments are responsible for the development, documentation, and implementation of privacy protections within their administrative data systems. Integrated administrative data from these public service areas are still affected by other relevant regulations, even though they do not fall under the regulations of HIPAA and FERPA.

Child Welfare Information

The Children's Bureau administers the Federal and State reporting systems that provide data to monitor and

improve child welfare outcomes. States are required by Federal law and regulation to collect information on children in foster care and on children who have been adopted under the auspices of a State child welfare agency. The Adoption and Foster Care Analysis and Reporting System (AFCARS) is a mandated reporting system designed to collect uniform, reliable information on children who are under the responsibility of the State title IV-B/IV-E agency for placement, care or supervision (45 CFR 1355.40). Federal legislation mandates that states receiving federal funding for child welfare services must demonstrate their capacity not only to collect reliable information, but also demonstrate their ability to protect the privacy of persons served by these systems. However, there are no distinct guidelines for states on how to implement this privacy protection, and each individual state must demonstrate to the federal government how it provides privacy protection.

Homelessness and Housing

Federal programs for Housing assistance and homeless shelter services also mandate the collection of administrative data on the clients who are served. The McKinney-Vento Homeless Assistance Act of 1987 (Public Law 100-77) is the first and only major federal legislative response to homelessness, providing for a range of services for homeless people (e.g., emergency shelters, transitional housing or job training). In 2004, the U.S. Department of Housing and Urban Development (HUD) published a Notice in the Federal Register calling for the development and implementation of computerized data collection activities, in order for jurisdictions to receive funding under the McKinney-Vento Homeless Assistance Act (National Law Center on Homelessness and Poverty, 2005). These systems, called Homelessness Management Information Systems (HMIS), were to be designed and implemented at the local level, to allow for each system to meet the local needs of the populations being served.

In addition to provisions for the type of data to be collected, the federal guidelines also provide standards for the privacy and security of personal information stored in an HMIS. These standards are based on recognized fair information practices (such as those embodied in the federal Privacy Act), and were developed after careful review of the HIPAA standards. In any case where an entity possesses information that can be considered protected health information as defined by HIPAA, the entity will be exempt from HMIS privacy and security rules and must adhere instead to HIPAA.

As under the Privacy Act, HIPAA, and FERPA, there are provisions within the HUD legislation allowing for disclosure of information about homeless individuals and families for the purposes of research. In this case, a homeless service provider can disclose information for academic research purposes when an individual or institution has a formal relationship with the service provider, as outlined in a formal written research agreement. This research agreement must spell out the rules and limitations for use of the information, provide for the return or disposal of data at the conclusion of the research, and restrict additional use or disclosure of data except as authorized under the original research agreement.

Juvenile Justice

The U.S. Department of Justice is responsible for providing services for delinquent children and youth. Between 1993 and 2000, more and more states enacted new legislation endorsing information sharing, in order to streamline services for these youth, for example, among juvenile justice agencies and school districts in response to increasing incidents of lethal school violence.

As state and local jurisdictions across the U.S. work to improve coordination of services for delinquent youth, the Office of Juvenile Justice and Delinquency Prevention (OJJDP) recently recognized juvenile information sharing (JIS) as an essential tool for decision-making (Mankey, Baca, Rondenell, Webb, & McHugh, 2006). Their report, Guidelines for Juvenile Information Sharing, provided a needed framework for the development of information sharing networks that includes a consideration of privacy and confidentiality. While it does not provide specific mandates or requirements, this report does refer to the federal Privacy Act as the gold standard for states to use when determining their procedures for information sharing.

Ethical Challenges

Federal and state laws are designed to protect individuals from misuse of personal information by providing strict guidelines for the collection, storage, and use of administrative data for research purposes. Though necessary, these laws are not sufficient to cover the full range of concerns related to the potential harm that can result to individuals and public agencies from unethical research. Careful attention to ethical challenges is necessary to genuinely fulfill the purpose of releasing protected administrative data to researchers. Fundamental to the ethical conduct of research is ensuring that the potential benefits of research with human subjects significantly exceed the risks.

Research with human subjects should be conducted in the best interest of the individual participants, safeguarded by their informed consent or the permission of officials charged with ensuring the confidentiality of their administrative records.

In addition to meeting legal requirements, researchers who wish to gain

access to integrated administrative data must navigate the explicit ethical challenges of Institutional Review Boards (IRBs) and the implicit challenges of establishing research partnerships with data sharing agencies.

Institutional Review Boards

Institutional Review Boards are mandated for any organization receiving federal funds that conducts research involving human subjects. The IRB is charged by the federal government to conduct formal ethical reviews of all research activities (45 CFR 46.102(a)). The level of review varies depending on the nature of the research project and the safeguards that are needed to minimize the risk associated with participation in the study. The Privacy Act defines three levels of review: full, expedited, and exempt. Full IRB reviews are required for any research in which the investigator will be collecting information directly from human subjects (e.g., clinical research to test the effectiveness of a given medication). This research presents the greatest level of potential risk, and therefore requires the greatest attention to ethical conduct. An expedited review can be considered in cases where the research proposal presents minimal risk to the participants, such as, during observational studies of students in educational settings.

A third category of IRB review is considered for research studies that propose to use existing sources of information, such as the use of administrative data systems. Federal regulations state that research involving the collection of existing data is exempt as long as the sources of information are publicly available or the information is de-identified (45 CFR 46.101(b)). Also exempt are research or demonstration projects that are conducted by or approved by department or agency heads, and are designed to examine the public benefit of service programs, procedures for obtaining services, possible changes in or alternatives to programs, or changes in methods of payment for services under those programs.

Research Partnerships

The implicit ethical challenges associated with entering research partnerships with public service agencies arise from the most fundamental ethical principles of research ethics: beneficence, respect for autonomy, and justice (Department of Health, Education, and Welfare, 1979). Beneficence calls for researchers to seek the best interest of the participant community. Respect for autonomy mandates responsiveness on the part of researchers to the informed choices of the participants. Justice prohibits any undue burden or hardship to participants as a result of their involvement as participants in
research. Adherence to these basic principles provides a foundation for participants and participating agencies to trust that the research will benefit all those involved and minimize risk to participants.

For public service agencies sharing sensitive administrative data with researchers, these ethical principles reflect three real fears. First, the agencies fear that if they share de-identified data, somehow the researchers will be able to re-identify individuals by using other sources of data. Ready access to the Internet and expanded computer capacities to search and link information fuel fears that personal information may be revealed. An entire separate body of research seeks to answer the question, "What is really de-identified?" and "How easy is it to re-identify?"

Second, they fear that findings from this research will be misinterpreted or disseminated in such a way as to unjustly portray the client population, the agency and service providers in a negative light. For example, a study showed evidence suggesting that a finding of a disproportionate representation of African American women in the crack (cocaine) use population could result in racial profiling at health care centers (see Leigh, 1998, for further discussion of this example). A third ethical consideration relates to a fear by participating agencies that the research will provide no concrete benefit to the agency to justify the expenditure of time and resources to make the data accessible. These service agencies are often serving complex client populations with insufficient resources. Therefore, they are reluctant to spend agency resources on projects with no tangible, real-time benefits.

Scientific Challenges

The ultimate usefulness of administrative data for research and planning purposes depends on the original data quality. Systems administrators must develop data acquisition, auditing, and linkage procedures that assure the data's integrity for research purposes. This is the science of integrated data systems. Computer scientists have developed methods for addressing many of these issues from a technical standpoint. These methods can range from complex real-time (or nearly real-time) relational databases with sophisticated record linkage systems, to more basic parallel archival processes with annual updates that are merged using stored record linkage procedures.

The scientific issues most commonly engaged by integrated data systems include issues of project administration and data integrity, as they relate to the science of enabling applied research activities.

Data Acquisition

How data is transferred from sharing to host agencies may vary depending on the data exchange agreements that are in place. The transfer process can simply involve the delivery of a CD, external hard drive or tape, while

some systems may use electronic transfer via File Transfer Protocol, a virtual private network, or through automated file transfer routines that send data from one system to another on a periodic (nightly or monthly) basis. Data sharing that takes place outside the firewall of a government-controlled infrastructure will need to include additional security protocols. Encryption and other data security mechanisms should be put in place to protect against unintentional disclosures of data due to loss or theft. Common strategies include using external hard drives with a built-in encryption system and biometric access keys, or running an automated file transfer routine on a designated secured computer. While the data acquisition process rarely affects the overall integrity of the data, the process must be designed with consideration for both the overall project management and partnering agencies.

Data Cleaning and Auditing

Database administrators typically perform basic review procedures that include cleaning and auditing the data. This includes the review of file specifications and record layouts that accompany data transfers. The host agency usually undertakes a review of any files received to make sure that the files match the specifications, and to assess their consistency with previous files from the data-sharing agency. Previous file versions from the sharing agency will often be stored by the host agency, and they will sometimes be merged with the received file so as to create an updated record. The database administrators must look for changes in the file layouts or other incompatibilities, since coding schemes and data fields may change—either being added or deleted—due to changes in policies or procedures of the data-sharing agency. If it is possible, the host agency must identify when these changes have occurred by providing an updated metadata file. A record of all changes and any variations associated with particular files is usually maintained as part of the metadata of any system.

Beyond the basic review of a transferred file, data-base administrators perform more detailed auditing of records to look for issues or problems with the data. Procedures can include variable-level auditing to look for out-of-range codes and for the frequency of missing data. Variables can be scored with a reliability measure, so that external requestors are aware of the reliability of a given variable. Common audit routines can measure the completeness of a given variable (degree of missing data), the accuracy (the proportion of valid codes), and coverage (gaps in time periods reported, or providers reporting, for example). When two data sources are available for a given measure, for example, diagnosis associated with a hospitalization, the two data sources can be compared to assess the degree of agreement between the two sources. Discordances may raise questions as to which source is considered more reliable, and may require further investigation.

Validity testing, another data auditing task, assures that data collected in a variable actually represents the phenomenon in question. In some cases this may involve manual checking of records from paper files against the electronic data. Due to its time consuming nature, this task may only be done on an annual or semi-annual basis. Since most agencies are not equipped to conduct such validity testing on a routine basis, IDS leadership may have to partner with data sharing agencies to periodically seek funding to accomplish these important audits.

A final advantage to the creation of an IDS is the support the host agency gives to the sharing agencies' efforts to improve data quality. Because many agencies are often too busy with business processes to assess their own data quality on a regular basis and because data quality is often contingent upon use (the most commonly used variables usually have higher reliability and validity), an external hosting partner who reviews the data can provide an opportunity for data improvement. The host agency can work with sharing agencies to develop internal procedures for improving data. Related research projects may also identify important gaps in information, and guide improvements in services. Of course, data sharing agencies also know their data best and can help the host agency to understand the nuances of their data in ways that may not be fully captured by metadata and record layouts.

Record Linkage

The critical advantage of IDS is in the process of record linkage. Record linkage refers to the joining or merging of data on the basis of common data fields, usually personal identifiers—commonly a name, birth date, and Social Security number. They may also include system-generated client identifying or tracking numbers, or mergings of multiple identifier fields into a "unique ID." In some cases, addresses may be used as a linkage field, particularly for projects where geographic location is central to the intended analysis.

A variety of tools are available to facilitate record linkage, and many organizations may have created their own methods for linking records. The key issue here is creating decision-making rules with parameters for determining what constitutes a matched (i.e., successfully linked) record. Keystroke errors, misspelled names, and the transposition of characters are just a few of the potential data problems that would reduce the number of correct matches. To minimize these false negatives, database administrators may perform the matching process using unique identi-

fiers created from parts of fields, such as the first two letters of last name and first name, month and year of birth. They may also use Soundex (or another phonetic spelling translation algorithm) as an alternative to exact name matches.

In general, two types of record linkage are possible: deterministic and probabilistic. Deterministic record linkage involves matching on the basis of an agreed upon set of data characters or strings of characters (with some allowance for missing data). Deterministic matching procedures are typically employed when users are most interested in reducing false positives, or the matching of records that do not belong together. Probabilistic matching procedures involve the use of algorithms that permit flexibility by weighing fields differently when assigning a match. This procedure is often used in large studies where false negative matches may be more of a concern, or when deterministic matching is not possible given gaps in common identifiers. Probabilistic methods can also identify potential matches prior to a deterministic matching procedure.

Link King, a public use data-matching software developed in part with support from the Substance Abuse and Mental Health Services Agency of the U.S. Department of Health and Human Services (Camelot Consulting, 2008; http://www.the-link-king.com), enables users to set probabilistic matching parameters across a variety of dimensions. Link King also supports deterministic matching. A particular strength of this software is its ability to generate a set of standardized statistics that measures the degree of certainty associated with a given match. Such statistics can aid researchers and consumers of research in identifying the criteria, stringency, and overall robustness of the match.
The science of record linkage continues to be advanced by statisticians and computer scientists. A bibliography of work in this area can be found at http://www.cs.utexas.edu/users/ml/riddle. Different users will have different purposes, and will want to be more or less sensitive to false negative or false positive errors. As communities develop these procedures and share their approaches, the field will see the development of consistent procedures for communicating matching protocols and of standards for assessing the quality of record linkage results.

Methodological Opportunities

The potential for scientific uses of an IDS goes beyond system-related issues. Given their population-based nature, epidemiological methods are often appropriate for grasping the basic incidence and prevalence of systems use, the relative risks for program entry, and outcomes for subpopulations.

Event History Analysis. The longitudinal nature of the data also provides opportunities for event history analysis to study patterns of program entry and exit, the duration of spells, and the hazard rates associated with subpopulations of program users, programs, and outcomes. Research on interventions that involve primary data collection and the tracking of a cohort of cases and controls can use an IDS to pull in relevant covariates (moderating and mediating variables) from the time period before the person was enrolled in the study and throughout their enrollment (instead of having to rely on self-reported data).

Time series analysis can be used to measure program utilization rates in the aggregate, to forecast program use in the future, or to measure participants' individual usage patterns, while controlling for program utilization variables from other systems.

Spatial analysis. The availability of geocoded data creates a variety of spatial analysis opportunities, including analyses of risk "hot spots" and the creation of aggregate measures of the social environment around individuals, for example, the count of truant or delinquent children in a given search radius around an individual's address.

Cost-accounting research based on imputed costs associated with various units of services consumption as well as using IDS data to generate cost and cost-offset data for benefit-cost and cost-effectiveness research are also among the analytic options created by an IDS.

In short, the large variety of analytic tools available to social scientists is readily applicable to IDS data, and, indeed, IDS data provide a very rich opportunity to explore questions from a multitude of approaches. The ongoing availability of these data provide opportunities for researchers to test models developed with one cohort on subsequent cohorts. This capacity and the relatively low cost of replication allows for refinement of models to best serve specific target populations. (See, e.g., Culhane & Metraux, 1997, for an IDS research agenda for homelessness and a discussion of related analytic strategies.)

Economic Challenges

Development Funding: Purposes

Any system has an initial purpose, even though it may evolve into serving a much broader set of needs. The system's initial purpose also usually reflects the source funding its development. Public agencies, especially executive branches of government, may see the value of integrated administrative data for managing large operations and multiple departments. Such data may play a role in government budget officials' evaluation of departmental requests for funds for various initiatives or in simulating criteria for

program eligibility. Private philanthropy may seed development of an IDS as a way of developing a research and evaluation capacity, in order to improve services for populations of interest (e.g., children exiting foster care) or for a given issue (e.g., prisoner reintegration). Indeed, researchers may initiate the development of a system as part of a research infrastructure and may seek public or private funding tailored to their research interests.

Operational Funding: Uses

Beyond the development costs associated with setting up an integrated data system, implementers will have an ongoing concern over the maintenance and sustainability of the system. Interest is usually highest among funders for the development period, but ensuring funds for maintenance and operations is often more difficult. Because of changing circumstances and competing priorities facing funders, such systems will often need to prove their worth. If the host is a public agency, political mandates and administrative uses may justify its ongoing support—at least for the duration of a given administration. If the host is a private agency, like a university, then operating support may need to be underwritten through research grants and contracts.

In either case, part of the planning for an integrated data system will need to include a business plan that offsets ongoing operating costs either through core support from local or state government, from private funders, through the conduct of contract or grant-funded research, or through the charge of usage fees. Indeed, it may well be in the interest of the enterprise to market uses of the data, both to data sharing agencies within government and to external research organizations, such as universities. As with all the challenges and partnership issues described above, each system will likely evolve its own funding solutions based on the purposes and functions of the system.

Cost-Efficiency of an IDS

Aside from these issues related to operating expenses, the relative cost-efficiency of an IDS for conducting research is worth noting here. It may be obvious that primary data collection efforts are much more time and resource intensive, at least for the researchers, than is the integration of administrative data sources. The costs of administrative data collection are underwritten as part of the business expenses of public agencies. The data are also generated on a continuous, real-time or nearly real-time basis. Data provide an alternative to periodic interview waves of study samples and self-reported data on program use, school attendance, health care use, etc.

The costs to researchers for accessing integrated data will vary based on the data sources and the number of hours required to process the request. An IDS can reduce the costs of individual data requests by maintaining procedures for cleaning and preparing all data for analysis as those datasets are obtained and stored. Otherwise, redundant costs may be incurred if files are prepared only in response to specific requests. A given IDS may also have operating support from other sources that can be used to offset the costs of data requests. However, it is more likely that external requests would be used to offset operating expenses. Typical database merges may range in cost from $20,000-$50,000 but could exceed that in the case of more complex files, such as, pulling data from Medicaid and other health claims across multiple years.

Given the temporal range of the data and the volume of potential observations, the IDS approach is significantly less costly and more efficient than primary data collection. However, trade-offs must be made. Consider, for instance, that a primary data collection project for a sample of 500 persons can take a year to enroll subjects, two years to track them prospectively (for 18-24 months), and one or two years to analyze the data. The costs for such a four- or five-year study would be $2-4 million, depending on the amount and type of data collection. In comparison, an IDS project can track thousands or tens of thousands of clients in a given intervention across multiple years and across multiple systems. Because the primary responsibilities of the researcher are the design and analysis components of the project, projects can be initiated and completed in a matter of months or perhaps one year or two in the case of more complicated projects. The costs of a typical project can be quite variable, but are likely to be less than $300,000 in more complex cases and substantially less in many cases. The comparatively low cost of an IDS-based research project makes more frequent and more time-sensitive study and analysis more feasible.

An IDS cannot substitute for primary data collection in certain domains. While an IDS may be used to track research participants in a randomized study, an IDS most often will be used for quasi-experiments, where control groups will be generated among non-participants. The risk of selection bias from a lack of randomization can be partially offset by enhanced opportunities for matching and for other statistical controls. Specifically, the large number of potential subjects in administrative data can provide for more comparable matching through the use of service histories. The large number of subjects also enables greater statistical control for pre-existing differences in the study groups. Nevertheless, where time or money are not an issue, an ideal approach would be to combine an experimental study with administrative data, where the power of randomization and primary data collection is combined with administrative data for tracking historical and prospective service use patterns.

A Survey of Exemplary Integrated Data Systems

The authors undertook a survey of project administrators working at existing integrated administrative data systems with the aim of learning how these IDS address the legal, ethical, scientific, and sustainability challenges outlined above.

Participating Sites

To identify exemplary IDS cases for inclusion in the survey, a key informant process was developed. Key informants with potential knowledge of integrated data sites were identified from among known administrators of existing systems, a federal human services research sponsor, and university researchers. All informants were then asked to identify any sites with which they might be familiar. This process generated a convenience sample of sites meeting the inclusion criteria. It did not include an exhaustive search for potential candidates. For systems to be included in the survey, they had to meet three criteria

- The IDS must contain data from multiple agencies.
- The IDS must have been developed as a general utility, rather than for a specific research project.
- The IDS must involve individual-level record linkage (aggregate level data integration was not sufficient).

Key informants identified eight exemplary sites as meeting the inclusion criteria, and all eight agreed to participate in the survey. They are

- State of Michigan
- State of South Carolina
- *County of Los Angeles*
- *Allegheny* County, PA
- University of South Florida (for the State of Florida)
- University of Pennsylvania (for the City of Philadelphia)
- University of Chicago (for the State of Illinois)
- Case Western Reserve University (for Cuyahoga County, OH)

Method

The survey, designed by the authors, was intended to collect data in the four areas outlined above—the legal, ethical, scientific and economic challenges that integrated data systems are likely to face. The survey also sought to identify the data sources for each system, and exemplary projects

Results

Legal Issues

Nature and Purpose of Legal Agreements	All of the data systems surveyed have in place legal agreements among the data sharing agencies. These agreements address common concerns among the data systems. • Legal agreements explicitly state that contributing agencies maintain control over data usage and release—despite being held by a host agency, data are regarded as the property of the contributing agency. • Three of the agencies surveyed also have specific stipulations regarding data security and confidentiality standards that must be observed by the host agency.
HIPAA Compliance	All of the data systems surveyed collect data that fall under the purview of HIPAA. All data systems conduct compliance audits: • 3 conduct internal and external audits. • 2 conduct internal audits only. • 3 conduct external audits only.
FERPA Compliance	5 of the 8 data systems surveyed collect data that are under the purview of FERPA. Of these, 3 conduct internal compliance audits.

Legal Agreements and Data Usage

Because data usage is controlled through legal agreements among data sharing agencies, the integrated data systems were surveyed to determine the method by which the proper usage of data was determined.

- Four of the data systems surveyed reported a formal committee comprised of the data sharing agencies that addressed data usage. In one of these data systems, the committee is formally created and governed by state statute.
- Four of the data systems surveyed did not have an individual or a formal committee to review data usage.

Review of Proposals and Projects

Of the 8 integrated data systems surveyed:

- 7 have a formal process for reviewing research proposals.
- 5 have a formalized process for reviewing completed research projects.
- 6 have a formal mechanism in place for sharing research finding with all of the contributing agencies.

Respondent Comments Regarding Legal Issues

All of the data systems surveyed noted similar legal hurdles and ongoing legal concerns. Some comments by respondents included

- "There were confidentiality requirements and legal barriers that prevented the sharing of client information among County agencies."
- "Government agencies have procrastinated on signing data sharing agreements."
- "It was difficult [for agency attorneys] to identify the specific terms of agreements [required for an MOU] because the governing legislation is different depending on the source of the data."
- "Some agencies were reluctant to share data if the law did not explicitly require data sharing."

that illustrate the value of these systems for research and policy analysis.

The survey took an online, structured-interview format. The research team contacted respondents about their interest in participating. All respondents agreed to participate, and were sent a link to the online survey. Respondents were asked to complete the questions to the survey to the best of their ability, recognizing that they might not have time for a detailed accounting for some items, such as exact number of records or variables. Respondents were given two weeks to complete the survey, and members of the research team followed up with respondents to ensure timely completion.

Analysis

Answers to the survey questions were assembled into a series of tables representing each survey category. Project staff analyzed the survey data and reviewed and checked it for accuracy. The data were used to create summary tables. Individual site responses are not provided here.

All the respondents cited specific data-sharing agreements that they have established with data-sharing agencies who set the policies for the use of data within their systems.

Agencies were most concerned about their compliance with the explicit federal privacy acts that govern data use—HIPAA and FERPA. The data systems surveyed collect large amounts of information which fall under the purview of HIPAA, including data from state departments of health and human services covering Medicaid/ Medicare and hospital payer data), community and state mental health agencies, care programs, and agencies providing substance abuse treatment. In addition, several data systems have collected FERPA-protected data from municipal school districts. As indicated in the box above, the use of internal and external audits in assuring compliance is a common practice among the integrated data systems surveyed. The collection of HIPAA and FERPA-covered data was sited as one impetus for creating formal data sharing agreements, suggesting that federal requirements may well have created a higher standard for formalized data protection protocols than would have otherwise existed.

Legal agreements between agencies are the foundation of integrated data systems. They address concerns over data sharing, use, and distribution, as well as the need for data security. In most cases, the contributing agencies retain full control over the use of their data, and, as some respondents have noted, data are treated "as still belonging to the agency from which it came." In almost all cases, the agreements take the form of memoranda of understanding (MOU) or memoranda of agreement (MOA). The period of renewal of MOUs and MOAs varies across data systems, and ranges from one to five years.

In addition to basic guidelines on confidentiality and data security, the agreements created by several of the data systems also stipulate the creation of a special committee to regulate data usage, create guidelines regarding how data requests are processed and how completed projects are reviewed, as well as lay down a mechanism through which research findings may be shared across the various data-contributing agencies. In four of the sites, research findings must be formally presented to data-providing agencies prior to dissemination. In the four cases where the sharing of findings is not required, two stipulated that participating agencies might at their discretion require the submission of findings for a specified period of review prior to dissemination.

A few respondents cited that the lack of an affirmative legislative mandate for data sharing created a perception among some potential data-sharing agencies that data sharing is not an important value. Further, it might disincline some agencies to commit the time and resources necessary to participate. In addition, individual agencies may have policies specifically governing data usage or may be under other state or federal regulations regarding confidentiality—like those covering earnings or income tax data—that limit their ability to participate in an integrated data system.

The common purpose for the creation of integrated data systems is to enable the simultaneous analysis of data from multiple agencies . However, data use is predicated on a foundation of ethics that protect data from misuse. Certainly, the respondents here identified confidentiality protections as foremost most among their concerns regarding the ethical treatment of data. In most cases, confidentiality is assured through multiple levels of protection. To begin with, some data systems have restricted data sharing only to other government organizations.

In a majority of cases, the sharing of data with researchers is mediated through a formal review process that ensures that the use of the data is in accordance with policy, and that any risks to confidentiality are mitigated as much as possible. All the respondents require formal agreements between the IDS and researcher when information is shared. In most of the systems surveyed, these agreements also require formal review and sharing of research findings. In one case, the process for acquiring data entails written permission from each data-sharing agency involved, rather than just blanket permission from the IDS.

Most respondents have technical safeguards for their data and train their personnel in the handling of confidential data. Data sharing with external partners is usually limited to de-identified data, although the survey results did not clarify the degree to which de-identification shielded dates of service (a limited dataset) or not. Even with rigorous technical and procedural safeguards, breaches of confidentiality are almost always possible.

As in any research involving human subjects, integrated data systems must weigh the risk associated with the research with the potential benefits. The administrators surveyed consistently reported that the power and importance of their respective IDS for informing social policy was profound, and, as such, its benefits to the population outweighed the associated risks. These benefits include better targeting of resources to needy or at risk populations, generating data that better inform social policy, and greater efficiency in the application of resources, resulting in notable budgetary savings.

The respondents varied significantly in how they address the various scientific issues associated with maintaining and using their integrated data systems. This variability may well reflect the relative maturity of the various systems; the largest contains 11,000,000 individuals and 32 years of data, and has been in existence for 20 years. There is an apparent relationship between the age of the systems and their size. Almost all the systems have been in existence for at least five years, with more than half over ten years old; collectively, the databases surveyed contain information on an average of 4.6 million individuals. Data acquisition appears to be a gen-

Ethical Issues

Data Sharing

Many of the systems surveyed have a formal system for the review of proposals and research results.

- All 8 provide data to governmental agencies.
- 5 of 8 provide data to private organizations.
- All 8 provide data to researchers.
- 7 of 8 require formal agreements between the IDS and the researcher in order to share data.
- 1 respondent stated that permission for data use must be individually requested from each contributing agency.

Research Review

Many of the systems surveyed have a formal system for the review of proposals and research results.

- 7 of 8 data systems surveyed have a formal system for the review of proposals and research results.
- 5 of the 8 have a formal process for the review of completed research:
  - Of these, 4 review all written materials.
  - 4 require a formal presentation of findings.
- 6 of the 8 data systems have a formal system for disseminating research findings among all of the contributing agencies.

Protection of Confidentiality

Concerns about confidentiality are at the heart of the agreements that allow for the construction of integrated data systems.

- One respondent noted that data sharing among agencies was only possible by using de-identified data.
- In at least 1 case personal identifiers are not stored with statistical data files.
- Data provided to researchers is de-identified.
- In all cases, agreements between data sharing agencies and between the IDS and the researcher have specific stipulations regarding confidentiality.

erally smooth process for most of the respondents, at least after the formal legal agreements regarding data sharing are established. One system did report difficulty getting all contributors to keep up with their data contributions, but others indicated that once a system for acquiring data was established and instituted via MOU/MOA, the difficulties encountered tended to be occasional, involving, for example,

Comments from Data System Administrators

- "With data contained in this system, we can better understand our demographics, the neighborhoods we live in, our socioeconomic levels, and the family structures that are so important to who we become as adults, the difficult circumstances and health problems that affect our lives and our independence and the cost to society for offering programs and services to those who are at risk."
- "They [policy makers] understand that the IDS is a unique resource that can provide them with information that they don't have and that their agency will likely never have the capacity to produce. Policymakers can learn the outcomes for the agency service recipients. They can learn the characteristic of children and families that come to their attention. They can use the IDB data to help identify their service population overlap with other agencies populations and develop a richer sense of need."
- "The integrated information on public services provided to indigent adults in the General Relief program allows to evaluate the services provided to this population service utilization patterns and the cost of the services provided. The information is being made available to policy makers to enhance the delivery of public services to indigent adults and to design new programs."

---

corrupted files or limited staff time. One site also observed that once the data are put into a "production environment," the ongoing maintenance effort is greatly reduced. Despite the large volumes of data processed by these databases on a yearly basis, few report major problems with data quality. Respondents' answers to questions about data quality focused primarily on the reliability of the data, and indicated they use automated routines to check for errors. The survey did not confirm that any of the sites had regular auditing of data for validity, a process that is inherently more complex.

One system reported that although "data is far from perfect…the originating agencies use this data in their day-to-day processes so the core data is fairly good." This is echoed by another respondent, who states that "most agencies have data quality standards in place so that the data coming to us are of good quality."

Another system has chosen to "work closely with the data experts from each department to clean the data from the participating agencies." Of all the systems surveyed, only one reported consistent problems with poor data quality from state agencies.

Both probabilistic and deterministic methods are reported as being used for record linkage, although only three sites explicitly referred to probabilistic methods. Most sites appear to use some version of a system-generated identifier or a concatenated identifier from among components of various identifying data as the basis for record linkage.

With one exception, these integrated data systems are funded by multiple sources. Most reported that their primary income sources were from state and local governments, and in one case, federal funding as well. One noteworthy exception received its entire budget from private sources, including a foundation grant and user fees. This mix of funding reflects the fact that most of the systems surveyed had their genesis in state and local governments, either through contracts or as a government-operated system. Some data systems report that although they are an expense for local and state governments, the data they provide allows for more efficient and cost-effective targeting of resources.

For half of the respondents, income also comes from data requests, based on the staff time and system overhead required to process requests. Finally, it is important to note that the amount of staff time dedicated to each of these integrated data systems was also somewhat variable. Some systems reported a fairly straightforward number of FTE's dedicated to the project, with an average of 2.75. One respondent reported up to twenty-five people as involved in processing data requests, maintaining the database, and inputting new data. The amount of time each person dedicated to this work was generally

Science Issues

Updates and Acquisition

The number of contributing agencies varies greatly across data systems.

- The smallest system surveyed has 7 data contributors; the largest system has 70 contributors.
- The average number of data contributors was 23.
- Across all data systems, most data sources are updated more than twice a year. The remaining data sources are updated at least annually.

Storage

The size and structure of the databases is highly variable.

- The largest system includes 50 terabytes of data.
- The smallest systems surveyed contained 7 years of data on 20,000 individuals and 200 variables.
- The largest system surveyed covered 10,000,000 individuals, with data spanning 35 years and 40,000 variables.
- Average number of years of data was 17, covering an average of 4.6 million individuals. (This calculation based on 6 of 7 databases. 1 respondent did not include this information.)
- 5 of the respondents utilize a centralized database, while 2 are distributed among multiple databases.

Linkage

There are three main linkage techniques utilized. (Three respondents did not reveal how they link records.)

- One data system uses personal identifiers to assign tracking numbers.
- Two utilize probabilistic record linkage.
- One uses probabilistic and deterministic methods.
- All systems are GIS enabled.

Comments from Data System Administrators

- *"We have found it useful to place data management in the hands of our statisticians; this provides them with detailed experience in the quality and quirks of the data, making their advice invaluable to data users."*
- *"Most agencies have data quality standards in place so that the data coming to us are of good quality. When we find quality issues, we have a natural communication link with the contributor and resolve issues together."*
- *"One of the better efforts we have implemented is the merging of geo-coding and GIS presentation with the data in the warehouse….It has allowed access to geographical queries that allow us to analyze data by distance without reference to maps. (An example of this is in the Family to Family program, a foster care improvement program sponsored by the Casey Foundation, where we are trying to ensure that kids are placed in their own neighborhoods.)"*

small, estimated at 10 to 25% of their effort (between 2.5 and 6.25 FTEs).

Exemplary Uses

The systems participating in this survey offer a wide range of examples of their use for informing public policy and for evaluating programs—too many to summarize here. Here are some striking illustrations of the kinds of initiatives that an IDS makes possible.

- In the County of Los Angeles, The Adult Linkages Project (ALP) is used to examine and track different cohorts of General Relief participants, examine their use of services across a broad spectrum of health, social and law enforcement services. The project has most recently been used to examine individuals identified as homeless, to determine how placement in housing has led to reduced use of services, and to forecast potential cost offsets to county government of further housing development and placements.

- At the University of Pennsylvania, investigators have conducted a series of cohort studies to identify early risks associated with poor academic and behavioral outcomes. Risks include early childhood poverty, homelessness, premature birth, neglect and abuse, out-of-home placement and lead exposure. Results have followed children through third grade to show the protective effects of formal early care and educational experiences on later educational success and school adjustment.

- At the University of Chicago, researchers at Chapin Hall have used their integrated data to examine the impact of residential placements on children in the child welfare system. The research has led to restructuring child placements to reduce unnecessary or ill-timed placements, and to improve child outcomes.

- In the State of Michigan, the state's data warehouse has enabled state government to roll out several statewide programs with greatly improved efficiency and to measure geographic variability in program enrollment and outcomes. Two notable initiatives have included the SHADoW homeless research database project and a "Family-to-Family" initiative to improve placement and management of the state's child welfare programs.

- At Case Western Reserve University, the integrated early child-

Economic Issues

Cost and Maintenance Budget

There is broad variability in the costs associated with the creation and maintenance of integrated data systems.

- Of the 8 data systems surveyed, 4 have a specific budget for the maintenance of data.
- The largest data system surveyed has an ongoing yearly budget of $ 1.2 million for staffing and maintenance.
- Lowest infrastructure cost reported was $50,000. Average infrastructure cost was 1.5 million, though most fell between $100 and $800 thousand dollars for the five systems that reported budget figures.

Funding Sources

7 of 8 systems reported funding sources. All of these obtained funds from multiple sources.

- 4 data systems report that 20-50% of their funding came from state government
- 1 data system reported that 55% of their funding came from federal sources.
- 3 report funding from local government

- 2 report funding from private sources, with 1 receiving its entire budget this way.
- 1 reported that 80% of its budget came from "other" sources.
- 2 of 7 reporting funding sources reported having regular, ongoing contributors of funding (presumably for operating support—in other words, not project-specific funding).

Usage Fees

4 of the 8 systems surveyed currently charge for data usage.

- Usage fees for the 4 were assessed by calculating staff time and infrastructure overhead needed to fulfill the requests.
- 1 system reported that they are considering adding fees for data requests.

Staffing

Staffing levels are assessed by calculating the number of full time equivalent staff (FTE) needed to meet annual operational activities. This includes both system maintenance and project-specific activities (these are not distinguished by the survey).

- On average, the systems surveyed utilized 2.75 FTEs.
- The largest system had 7 dedicated FTEs.
- One system noted that there are perhaps 25 people working on the database, but that it may account for only 10-25% of their individual efforts.

Comments from Data System Administrators

- *"Since our statisticians do the data management as part of their functions, there is no specific budget for data management. This system has been largely built and is maintained by funding from data partners who see the value of the system."*
- *"Most of our projects have been initiated by agency management which has provided initial funding. To ensure that this continues we publish a regular report on what has been accomplished to remind the funders of what they are getting for their money."*
- *"We take a small part of research project funding for data management."*

hood data system allowed researchers in Cuyahoga County to determine the degree to which investments in early childhood programs were reaching all newborns, toddlers and preschool children, whether the timing was optimal, and whether the intensity of participation met the levels required. Gaps were identified. New measures initiated to actually bring proven programs to a significant proportion of the population and to estimate the benefit of these investments in the region's children.

- The University of South Florida has used its integrated database to facilitate several service system reforms for people with mental illness. Recent projects include a cost-benefit study of a

medication for Medicaid recipients with Alzheimer's disease. Other recent research has included studies of the effectiveness of specialized therapeutic foster care, the effectiveness of children's mental health services, and patterns of juvenile justice system involvement of youth with mental disorders.

- South Carolina has worked with state agency partners to create several analytic "cubes" so that agencies can drill down into pre-aggregated data to very fine levels of detail. For example, one education-oriented project enables policymakers to study the relationship between poverty, health conditions, crime, mental illness and success in school. The technology permits the cube user to select an analytic cell, to drill down to de-identified data, and see a full client history. Other projects in South Carolina, including the web-based electronic medical record, enable users to access, pending patient consent, identified data.

Recommendations

Having developed without any national program mandating practice-based standards and guidelines, the integrated data systems surveyed here give a picture of a diverse set of IDS innovators. They represent a continuum from IDS located within government to those created by research universities. Systems within government have the highest level of official public support. These systems work under direct administration by state budget or executive branch IT offices. Independent, university-based systems primarily use their integrated data capacity to obtain private and public funding for their faculty's research.

Some hybrid forms lie between public and private poles. These include efforts led by county governments that also work with external research organizations. One university that hosts various county data, makes them available to public and private organizations with authorized research projects. Regardless of their distinctive context, each IDS has had to address a common set of legal, ethical, scientific and economic challenges. From their collective experience, some conclusions can be drawn regarding best practices, and some recommendations offered for how other communities (states, localities, universities) and public policy stakeholders (federal agencies, foundations, policy research organizations) can consider the potential value of integrated administrative data systems for improving the effectiveness and efficiency of their public service systems.

Creating a Professional Learning Community

During survey-related conversations, respondents expressed a need to create a community of other experts like themselves for sharing information, organizational strategies and technology associated with the administration of an IDS. The range of technological sophistication among the respondent sites varied widely. In general, the state government sites had the largest and most sophisticated systems, likely reflecting both the tenure of their systems and their position under the aegis of the executive offices of state government. Local governments occupy a middle ground, and universities had the simplest platforms, with no real-time data integration.

Regardless of these variations, the sites have much to learn from each other regarding all of the aspects of the IDS administration, from sharing templates for legal agreements, to the design of policies and organizational structures for processing research requests. Some sites had more experience than others in the creation or use of tools for querying the data, or both, including the use of pre-aggregated data cubes, or in using GIS. Other sites had more experience with sophisticated data linkage algorithms.

Regardless, all sites expressed interest in sharing technology, organization and financial operations, and in learning how to engage external researchers to maximize use. Thus, one important recommendation to come from this review is that a professional learning community should be formed among the sites to facilitate this exchange of knowledge. Such a group might also be able to provide technical assistance to entities which are considering the implementation of an IDS.

Establishing a Partnership Model: Optimizing Roles and Responsibilities

The variability in how the sites are organized and financed, as well as their robustness with regard to data and to use, suggests that it is possible to imagine a hybrid model of the various approaches identified here that draws on the strengths of each. The model partnership would optimize the most appropriate roles of the potential partners and maximize the use of the data infrastructure for policy analysis and planning. While no single site embodied perfectly this imagined hybrid model, a few of the various organizational approaches include most of its components, and suggest that such a model partnership is indeed possible and desirable, with appropriate partnership roles for government, universities, and funders.

Government

First, the survey results offer convincing evidence that government is in the strongest position to act as the lead agency with regard to the archiving of administrative data. Government has the authority to store vast quantities of data, as part of its responsibility for administering various public programs. While the clear authority for a single government agency to store multiple agencies' data may not always exist, state and local governments have shown that they can negotiate that authority under the appropriate legal agreements. Commensurate with this authority, government has the resources to store the data that are likely to be involved in any integrated data system.

Some universities have similarly shown that they are capable of being the host entity, and they may well be the appropriate choice in a given a locality, especially where agencies prefer a neutral third-party repository. For example, some agencies may feel that a neutral third party is less likely to inadvertently use linked data for program operations or client contact, since the third party may not deal directly with clients. Or, in cases where a city or local school district and county may have a conflict, a city agency or school district may be more willing to share data with a neutral third party than with the county. For now, however, the university efforts appear to be much less robust in terms of the number of participating agencies, their storage capacity, and their use of the most sophisticated computer science for data integration.

Similarly, the public agencies that host IDS have more robust financing, with ongoing commitments from government sources, whereas the university-led efforts are more likely to rely on periodic research contracts and private funds with no clear provision for infrastructure capacity building. These financing differences are reflected in staffing levels. While governments thus appear to be the most well-equipped to sustain these efforts, it is worth noting that to the extent that an IDS is part of and identified with the initiative of a given administration, there is a risk that the system will lose support with a change in political leadership. This would seem to speak to the value of a neutral third party, such as a university or research institute; however, even those arrangements can be changed with a change in political leadership. Shielding an IDS from political shifts, perhaps by housing it in an executive or legislative budget analysis office, could offer a long-term advantage.

Thus, with respect to two of the critical domains for an IDS, its legal authority and its economic sustainability, government appears to be the optimal lead partner. We would recommend that future efforts of this sort explore such a solution wherever possible. We also recognize that a neutral third party approach is possible and sometimes preferable, especially where numerous local authorities (for example, several school districts or several police departments in a given county) are unwilling to share data with a single government entity (for instance, the county). In such cases, the contributing public agencies will certainly need to extend their legal authority to the neutral third party. They should also consider how to provide financing for the maintenance of the integrated data infrastructure. Without that, nongovernmental enterprises are at a serious disadvantage in terms of their long-term economic sustainability.

Research Universities and Private Research Organizations

Independent researchers from either research universities or private research organizations have an important role to play in the effective use of integrated data systems to maximize public policy reform. Just as governments are uniquely situated with respect to the legal and economic aspects of an IDS, researchers are well positioned as partners to lead with respect to the science and use of an IDS. They bring a particular content expertise to the partnership in the social and health policy areas in which they conduct their research. This means they are aware of the latest research literature and state-of-the art research methods to address complex policy problems. University and other nonpartisan researchers are more likely to be independent participants in the local policy environment, and, as such, have greater autonomy and flexibility to access private and public research funds in support of support their work than staff members in local governmental agencies. All this makes researchers ideal users of an IDS and integral partners in a model implementation.

Partnership with a university or any external organization also brings with it some risks. Several of the government initiatives surveyed had undertaken relatively little partnership with academic organizations, possibly because of some of the perceived risks of opening access to the IDS to nongovernmental agencies. Those risks include not only obvious concerns with data security and confidentiality, but also the ethical risks associated with an external entity having access to governmental information, including concerns that the research findings would provide no direct benefit to the data sharing agencies or that the findings might be used unjustly to critique existing policy. Several of the respondent sites have been able to address these concerns through clear procedures for vetting proposals and research results and by affirming the

right of data sharing agencies to veto use of their data or to review and comment on findings.

To the extent that these ethical uses of data can be assured in a given partnership model, more jurisdictions may be willing to engage in these partnerships. Given the right organization and functioning of an oversight board consisting of researchers and representatives of data sharing agencies, the benefits of universities' participation can be structured to outweigh the risks, and succeed in leveraging the resources of both academic institutions and the IDS infrastructure for the improvement of public policy.

We think that the benefits of including universities as partners outweigh the risks. We would recommend that academic partners be included in the development of a system from its beginning. One possible mechanism is to include academic researchers on the IDS oversight board or on a specific "research review" board. The research review board could act as the scientific reviewer for projects proposed and completed to assure the academic integrity of the work being done. Inclusion of academic partners can also help to get some inaugural projects undertaken to demonstrate the value of the IDS to funders, data sharing agencies and other stakeholders. Creating a strong association with academic researchers from the outset can not only assure that the IDS is not an insular resource, serving only the more mundane management needs of government, but that it is an actively used resource for policy reform.

Funders

Finally, foundations and other research funders (such as federal research agencies) should be considered integral partners in an effective model. Foundations bring their own distinctive purposes and funds to support these purposes. Their missions are typically associated with assisting a given locality, a special population, or a certain interest area. As such, they can be important brokers in a community, and between government and external researchers. As independent partners, foundations or other research funders can use external funding to help to establish an IDS, and to create specific processes through which an IDS can be accessed in an ethical manner for research and evaluation projects.

Funders can establish conditions for funding that could both protect the ethical use of data by researchers, and the maintenance of transparent procedures for data access and research dissemination. National and local funders could partner in bringing stakeholders together to help establish appropriate protocols for an IDS or for research using an IDS, as well as to fund grant competitions for various research priorities among jurisdictions with an IDS. In any case, as shown in some of the current survey results, foundations can play an integral role in promoting an IDS and in bringing academic and government partners together around issues of common cause.

While no single survey site perfectly embodied the ideal model described here, each of them could benefit from greater engagement of the various partners. Our recommended partnership model leverages the appropriate roles and responsibilities of the respective partners for the maximum use and benefit of an IDS. As communities contemplate the creation of such a system going forward, they may consider this model for its applicability, or for how their local solution can benefit further from these roles and responsibilities under whatever model makes the most sense for that community.

Leveraging Capacity for Knowledge Development

Although our survey was not exhaustive, it included a variety of IDS sites. They range across states from different regions of the US, counties of vastly different size, and university-based efforts in large and medium-sized cities. These sites, as a whole, have a tremendous capacity to track large cohorts of individuals through multiple systems, across relatively long periods of time, and at a modest cost. However, up to now, these sites have not worked together to capitalize upon this capacity to engage in a collaborative multi-site study of a problem or population of mutual interest. Thus, in addition to creating a learning community of the administrators of these IDS, research funders should consider the possibilities created by the existence of these sites for pursuing greater understanding in various social problem and policy areas.

One way to enhance this potential would be for large national funders, such as foundations or federal research agencies, to help sustain the development of these nascent operations by funding specific research projects. An RFP could be targeted specifically to sites with an IDS or sites with an IDS could be funded to work together under a common research design. Some examples of cooperative research subjects could include:

- Assessing the impact of public or assisted housing on the long-term outcomes of residents, including school truancy, graduation rates, or negative social outcomes, such as, child abuse and neglect or homelessness.

- Examining the impact of early childhood program participation on school achievement and delinquency. Other studies could

examine the impact of adults' access to community-based health or mental health services on their employment patterns.

- Looking at a host of issues related to incarceration and prisoner-reentry, including the impact of incarceration on convicts' children and the impact of reentry programs on children's school success.

A variety of issue areas could be explored, either on a competitive basis for the communities with an IDS, or on a collaborative basis, exploiting the potential of these IDS sites to answer important issues using a standard methodology.

Fostering Replications

IDS is a field waiting to blossom, and national leadership is needed to actualize this vast, unrealized potential. The benefit of an IDS is clear—enabling communities to more carefully examine need and the utilization of public resources, thereby helping them to improve and maximize the use of those resources to achieve the best possible outcomes for many vulnerable populations. The opportunity is clearly present: every government agency collects data and usually does so as part of its existing business practices. Connecting these data and building a research capacity opens whole new areas of policy analysis and reform. Furthermore, as the examples from our survey show, the requirements of creating legal agreements and authority, partnerships among stakeholders, and identifying basic infrastructure support are surmountable, particularly relative to the gains to be achieved.

Consistent with the partnership model described here, one possibility for envisioning more widespread adoption of integrated administrative data systems is through the collaboration of funders, government and universities. National funders, in particular, could use their influence and resources to seed these collaborations in multiple sites throughout the country. In general, we see three stages to the development of an IDS.

- First is the collaboration and planning phase, which includes identification of the relevant stakeholders in a community and their agreements to participate. Successful completion of the planning process could be demonstrated by signed memoranda of understanding by the partners, committing the partners to the storage of the key data sources, and to the policies for the ethical use by both government and external research entities.

- The second phase is demonstration. In this phase, success is realized through the functioning of a data use oversight body, the actual storage of data and establishment of data updating procedures, and the use of the IDS with the successful completion of several research projects.

- The final phase is institutionalization, in which a regular flow of projects is established, and regular sources of funding are identified for maintenance and for research. A replication strategy mindful of these stages could help to build the nation's capacity for this important work.

Research Translation and Policy Reform: Some Demand Considerations

The power and utility of the IDS are only as good as the ability of policymakers to translate research results into actionable policy decisions. The capacity for such translational use should not be presumed. Individual agencies have widely varying capacity for using data to shape decision-making. Some agency executives are more data-savvy than others, and some agencies have more or less of an established culture for using data to inform policy and practice. This is not a new problem, and it is not unique to the use of IDS outputs. However, the analytic and translational capacity of public agencies does bear consideration if we are to make progress based on potential investments in data integration and research. Therefore, part of the effective implementation of an IDS may well be establishing a plan for cultivating the intelligent use and translation of research-based reform strategies within and across government agencies. Several efforts might be considered for enhancing that capacity.

One strategy for creating greater and more effective use of IDS and IDS-related results could be the training of a cadre of users through a special program. Ideally, Master of Public Administration (MPA) and Master of Public Policy (MPP) curricula could be used to develop this capacity, but it is also possible that a special certification program could be used to supplement these curricula, training MPA and MPP students in the use of these systems. Especially as these systems adopt more sophisticated real-time analytic and querying tools, program analysts may themselves become system operators. A special training program could help to develop persons who understand both the nature of the social policy areas in which they work and how to manipulate and interpret the multisystem data they can access to inform that work.

Perhaps a specific subspecialty of such a program analyst is the person working for the executive or legislative budget offices of state and local government. Educating state and local

budget officials in the potential of IDS may lead to both greater interest in their creation and maintenance and in their use to inform policy decision-making. Thus, it may be useful to consider some specific marketing of these systems and their potential uses to these people and their professional associations.

Finally, knowledge of the IDS approach and use of its outputs must somehow reach the deciders. State and county executives (including department heads) need to be educated regarding the value of these data for informing their prioritization of initiatives. As with the benefits and costs of the broad variety of program approaches, executives need to be aware that this capacity is possible within their jurisdictions. They can cultivate expectations that agencies will regularly report not only on the use of their programs, but on the indirect or secondary effects of their programs on other agencies' programs and costs. Especially as relates to "high needs" cases, who use the bulk of agency resources and are usually multisystem service users, government executives can use an IDS to envision the benefits of a more integrated service delivery system that maximizes the use of resources across agencies.

Only the executive branch sees across agencies and is concerned not only with funding in a particular department, but across departments; it therefore has the greatest need for agencies to be accountable for how they interrelate to leverage the resources of other agencies to solve complex problems. Making executives aware of these opportunities is perhaps best done in the context of specific problem areas. One possibility is to create various executive forums on specific topic areas , for example, children's mental health or prisoner reentry, where multi-agency data and research can be presented to illustrate impacts and best practices. These forums can communicate not only the content of interest, but also the value and exportability of the approach.

Concluding Thoughts

The central information problem in government and policymaking is not that there aren't enough data to answer a given question, but that the data (and the programs and resources) are partitioned in various departments. The executives of these departments may meet occasionally in cabinet meetings, but rarely do the program managers, policy analysts and the research and evaluation staffs of the agencies ever encounter each other. This insularity has given rise to the well-worn metaphor of agency silos and to the frequent complaint that the government systems don't talk to each other. The promise of an IDS is that it can move us beyond the paralysis of agency insularity. Before we can have more effective and more efficient policy-making, we have to establish a capacity for interagency dialogue. The medium for that dialogue is inter-agency data integration.

We also can't stop at the technology solution, and think that we have accomplished everything. If the data can't be translated into quality information, and if the information can't also be translated into actionable tasks, we have created just another silo. Real use of the IDS capacity will depend on partnerships for quality information and systems reform. Partnerships are needed that go beyond government. They link the research community, reform advocates and private sector interests, including business leaders and foundations. Together they can cultivate the use of data for the real substance of reform.

The promise of the IDS is not just technological. It lies in the promise that by building the capacity, we can also build the partnerships that will be the foundation of newly integrated systems of decision-making, planning and reform, that will advance our ability to address the many complex social issues that lie before us.

References

Camelot Consulting (2008). Link-King software. http://www.the-link-king.com/index.html (accessed August 8, 2008).

Culhane, DP & Metraux, S (1997). *Where to from here? A police research agenda based on the analysis of administrative data*. In D. Culhane and S. Hornburg (eds), *Understanding Homelessness: New Policy and Research Perspectives*. Washington: Fannie Mae. (http://works.bepress.com/dennis_culhane/8/)

Family Educational Rights and Privacy Act (FERPA), 20 U.S.C. § 1232g (1974).

Health Insurance Portability and Accountability Act (HIPAA), 45 C.F.R. § 160 (1996).

Leigh, W.A. (1998). *Participant protection with the use of records: Ethical issues and recommendations.* Ethics & Behavior, 8(4), 305-319.

Mankey, J., Baca, P., Rondenell, S., Webb, M., and McHugh, D. (2006) *Guidelines for Juvenile Information Sharing*. U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention (NJC 215786).

National Law Center on Homelessness and Poverty (2005). McKinney-Vento Homeless Assistance Act (PL100-77). Available at www.nlchp.org.

Address correspondence to:
Dennis P. Culhane
3701 Locust Walk
Philadelphia, PA 19104
culhane@upenn.edu