

Using Maps to Promote Health Equity

This report is one in a series of papers on best practices for using maps to promote health equity. Commissioned by The Opportunity Agenda, in partnership with the Health Policy Institute at the Joint Center for Political and Economic Studies, this project was made possible by The California Endowment. The complete volume of research and case studies is available on-line at: <http://www.opportunityagenda.org/mapping>.

The Numbers Behind the Maps

Michael L. Rodrian

Jim Watkins, CPA, MPPA

June 2009

Contents

Introduction	1
Methodology	1
Background	2
California State Vital Records Birth Data	3
California State Vital Records Death Data	4
Department of Health Care Services Medi-Cal Data	4
Office of Statewide Health Planning and Development Hospital Discharge Data	6
Environment	6
Limited Policy Direction	9
Limited Resource Allocation	17
Conclusion	19
Recommendation 1: Create an Organization With an Express Mission to Coordinate Data Releases Across Multiple Departments and Datasets.	19
Recommendation 2: Develop a Funding Source for the Establishment and Operation of This Organization.	20
Recommendation 3: Establish the Capacity to Match Datasets From Different Departments.	21
Recommendation 4: Develop, Maintain, and Operate Automated Systems to De-identify Data.	21

The Numbers Behind the Maps

By: Michael L. Rodrian, Jim Watkins, CPA, MPPA

January 2009

Introduction

The primary goal of this paper is twofold: to examine the obstacles to democratizing health care datasets collected by government entities and to recommend policy and/or organizational changes that will help achieve a democratized government health care dataset. In this paper, we will discuss the difficulties individuals, institutions, and others encounter when trying to access neighborhood-level data to measure health status and effect positive social change. While there is little doubt that the rich repositories of health care data maintained by governmental entities can serve the public good, it is difficult for most researchers and communities to make use of this data. It is even difficult for governmental agencies to share data among their sister agencies, many of which serve the same population. Some postulate this stems from the bureaucrat's desire to maintain ownership and limit access. Others have concluded that the bureaucracy is simply unable to deliver such data and is paralyzed by inefficiencies and untrained staff. And still others have simply laid the blame on the post-HIPAA environment. While evidence certainly exists to support aspects of each of these theories, we find that much of the resistance to releasing data stems from sensational news articles and political responses to data breaches that create a risk-adverse environment, limited and conflicting policy direction, a lack of a formal organization with an express mission and responsibility to coordinate issues between and among departments and datasets, and limited resource allocation.

Methodology

Through a case study analysis, the authors consider the environmental constraints of sharing governmental data with communities, researchers, and others who are seeking it in order to build community capacity to effect social change. The primary focus is on the governmental organization, looking through the prism of the governmental data steward. We examine the challenges faced when attempting to democratize data and the perceptions of those who are trying to use this important dataset to effect change. We also reference those who have used and/or requested government data to measure outcomes, perform research, and influence public policy and consider their experiences and outcomes relative to the government data request. We look to these data requesters for their perceptions and an understanding of their needs relative to timing, cost, and other issues of concern. In addition, we consider such data requests from the governmental data steward's perspective. With what issues must the data steward contend, and what obstacles must he or she overcome? Our focus is on four datasets maintained by three California government departments: (a) birth data and death data from the California Department of Public Health, Center for Health Statistics (CDPH-CHS); (b) Medi-Cal (Medicaid) data from the Department of Health Care Services (CDHCS); and (c) hospital discharge data from the Office of Statewide Health Planning and Development (OSHPD). Special attention is paid to geographically referenced data and the challenging issues behind its use in mapping community health data.

Background

To understand the concept of democratizing data, a definition is in order. While a number of definitions exist, the one that highlights the important aspects of this concept as used in this paper is derived from the National Neighborhood Indicators Partnership (NNIP). The NNIP was formed in 1996 to work with local data intermediaries¹ to provide direction and information that community leaders and others could use to effect change. In democratizing information, NNIP partners “facilitate the direct practical use of data by city and community leaders, rather than only preparing independent research reports on their own. And all have adopted as a primary purpose using information to build the capacities of institutions and residents in distressed urban neighborhoods.”²

Like the main concept expressed in the NNIP definition for democratizing data, our focus is on facilitating the direct practical use of governmental health care data by community leaders, local public health departments, researchers, and others attempting to effect social change or add to a body of research.

Governments in the United States collect large quantities of data in general, and the health arena is no exception. California, with its large and varied population and geography, provides a good cross section of the types of governmentally collected health data. Not surprisingly, the largest of these datasets—the Medi-Cal dataset—was created to collect fiscal information about health program costs. The principal health-related data gathered by the state of California includes procedure-based claims, eligibility information, and provider information from the state’s Medi-Cal program; discharge diagnoses and procedures rendered for each inpatient hospitalization and emergency room visit provided to OSHPD; and basic medical information about mothers and their newborns for all births in California and manner and medical cause for all deaths in California from the state’s vital records system.

Each of these databases is large, with hundreds of thousands or millions of annual records. The vital records and the hospital discharge datasets are also comprehensive, containing records of all events within the state. The Medicaid program in California covers a comprehensive set of medical services and compiles individual and procedure-level administrative data on roughly 8.5 million³ Medi-Cal beneficiaries each year.

Governmental databases tend to be large, comprehensive, and relatively stable over time. This minimizes bias, adds statistical power, helps with longitudinal studies, and allows for community

¹ These data intermediaries are nonprofit organizations that collect and compile datasets that can be used by community organizations, cities, and others to effect social change. They provide not only access to the data but also support of its use. For a discussion of their activities and objectives, see Elizabeth H. Guernsey and Kathryn L. S. Pettit, *NNIP Data Inventory 2007: A Picture of Local Data Collection Across the Country*, National Neighborhood Indicators Partnership, The Urban Institute, December 2007.

² Guernsey and Pettit, *NNIP Data Inventory 2007*.

³ This represents the total number of Medi-Cal beneficiaries with at least one month of eligibility throughout the calendar year of 2007.

comparisons. California's large and diverse population adds another element to governmental health care data that many states cannot offer: a statistically robust number of events to study. The large number of events allows researchers to study certain subpopulations, such as different racial and ethnic groups, as well as specific conditions that do not occur frequently among populations. Additionally, the charges for retrieving data are relatively minor, given the amount of data that can be obtained. This is especially true when the full costs of gathering and compiling your own primary dataset are considered. Further, governmental databases tend to be trusted by researchers and the general public. Some recent examples of targeted focus studies using the databases studied in this paper include:

- Health status analyses done by researchers, local health departments, and community leaders, such as looking at asthma in children, by neighborhood, with special attention paid to local issues such as environmental factors and adequacy of health care access in these neighborhoods
- Differential rates in cesarean sections performed on mothers in specific geographic areas and by hospital
- Medical treatment required as a result of violence
- Injuries, including those caused by auto accidents
- Health care costs and health insurance coverage by neighborhood
- Health services to the elderly
- Hospitalizations associated with ambulatory care sensitive conditions by payer source
- Cost-effectiveness of Medi-Cal managed care

In this paper, we reference specifically four datasets maintained by three California government departments: (a) birth data and death data from the Department of Public Health, Center for Health Statistics (CDPH-CHS); (b) Medi-Cal (Medicaid) data from the Department of Health Care Services (CDHCS); and (c) hospital discharge data from the Office of Statewide Health Planning and Development (OSHPD). These datasets are used to highlight specific points throughout this paper. We draw on our experience using these files as well as that of researchers who have used them. Although these departments maintain public use (i.e., de-identified) files, they are of limited use for community and local public health mapping. Therefore, our discussion focuses on the confidential datasets maintained by these departments. Three of these datasets are common to every state in the United States, with only minor variations. The hospital patient discharge dataset, with some variation, is present in about one half of the states.

California State Vital Records Birth Data

The CDPH-CHS collects data on all births that occur in the state. Birth data includes vital registration data (often termed "demographic data") as well as medical information, especially regarding complications to either mother or child. While comprehensive, including all births in the state, these data capture only the health issues surrounding the birth. The collection system does not capture hospitalizations immediately preceding the birth event or postpartum care to either mother or child. No information about later outcomes is available except for infant mortality, when the birth file is matched with information for deaths that occur in California. While some out-of-state deaths are also linked, the out-of-state information is considered incomplete. Despite its limited medical scope, the birth file remains a good source for capturing differential medical outcomes related to race and ethnicity, age, and other demographic

information about the parents, especially the mother. The birth file contains the mother's address, so data can be arrayed geographically. The data are reasonably current with the annual file, usually produced within six months of the end of each calendar year. The addition of a question about smoking history for the mother and the recent addition of the mother's weight to the birth file will yield additional public health information and benefit community health activities. Examples of customary uses of the birth file include studies of maternal mortality, low birth weight and premature birth patterns, prenatal care and its effect on infant and maternal health, and differences observed in post-hospitalization survival rates and overall differences in death rates by age, ethnicity, and region. This dataset includes elements such as the age of the child's mother, street address, city, county, state, ZIP code, gender, ethnicity, and Social Security Number. Additionally, the file contains the name of the child's father, and the date and time of birth. Substantial medical information about the mother, the gestation period, the delivery method and circumstances, prenatal care, and medical information about the child is also included.

California State Vital Records Death Data

The CDPH-CHS collects data on all deaths that occur in the state and includes both the manner of death and its medical cause. The decedent's last residence address, as well as the place of death, if different, is provided, yielding good mapping utility. Since the data are collected continuously, areas where chronic disease prevails can be monitored and targeted for outreach with better precision. Death data can also be linked with birth information for people who were born and died in the same state; this has been a primary source of health trend information for many years. To this end, the CDPH creates an annual death-to-birth linked file. Work has been ongoing to create an automated information exchange system, but to date it is both voluntary and only partially automated. Principal barriers to achieving an interstate automated system include financing for development and maintenance and difficulty in negotiating data-sharing policies given the differences in state laws. The California death dataset includes elements such as the decedent's street address, city, county, state, ZIP code, gender, ethnicity, Social Security Number, and mother's and father's names, as well as information on the date, time, cause, and manner of death.

Department of Health Care Services Medi-Cal Data

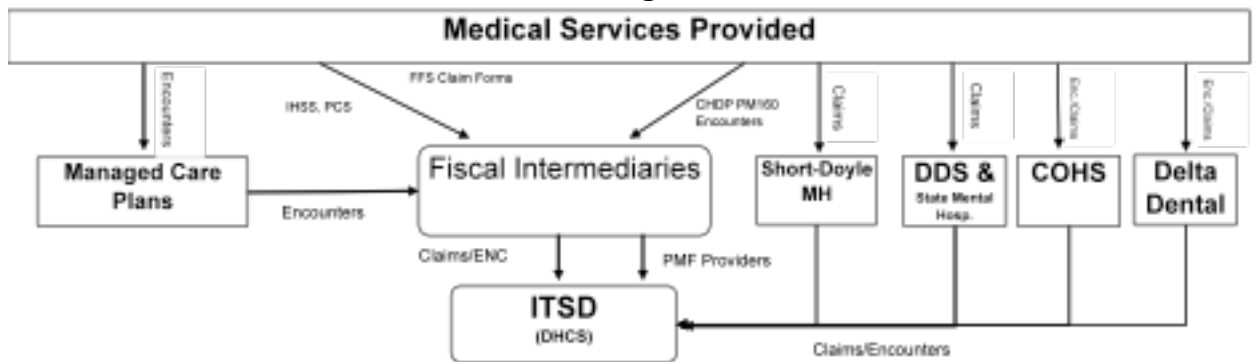
The Medi-Cal Program serves the largest Medicaid population in the country, financing health care services for more than 6.5 million beneficiaries monthly. Health care expenditures are expected to exceed \$34 billion during the 2007–08 fiscal year. More than 57,000 unique provider numbers, including 21,405 physician and physician group identification numbers, are reimbursed by the program each year, including 21,405 physician and physician group identification numbers. It processes 8.2 million fee-for-service (FFS) claims and pays more than \$2.7 billion dollars each month on behalf of 3 million users.⁴ In addition, DHCS collects and compiles

⁴ Numbers were derived from California DHCS, Medical Care Statistics Section, <http://www.dhcs.ca.gov/dataandstats/statistics/Pages/MCSSHomePage.aspx>.

encounter data on behalf of the 3.3 million Medi-Cal beneficiaries enrolled in contracted managed care health plans.⁵

The Medi-Cal dataset includes administrative paid claims data generated by programs administered by the California DHCS as well as other programs, such as the California Department of Mental Health, the California Department of Social Services, and the California Department of Alcohol and Drugs Programs. In addition, the Medi-Cal Program contracts with organizations such as managed care plans, and as part of these contracts, managed care plans must submit encounter data to the DHCS (see Figure 1).

Figure 1: Flow of Paid Claims Into the Medi-Cal Program



Source: Author prepared based on information compiled by DHCS, MCSS.

The administrative claims dataset includes individual and procedure-level data for almost all health care services that Medi-Cal-eligible beneficiaries receive. Included are pharmacy, long-term care, hospital inpatient, outpatient, medical/allied, and vision services and “crossover” claims for beneficiaries with both Medicare and Medi-Cal eligibility. The paid claims dataset includes data fields such as beneficiary unique identification number, county, aid code, name, gender, and race and provider ZIP code, name, unique identification number, county, and specialty. DHCS also maintains a master file containing information related to participating Medi-Cal providers, including data elements such as provider name, address, ZIP code, city, billing address, service address, and unique identifier. These two files may be and are routinely linked.

In addition to the claims and provider data, Medi-Cal maintains a complete eligibility history for each beneficiary. This dataset includes elements such as beneficiary name, street address, city, state, county, ZIP code, gender, ethnicity, and language. Again, this file can be linked to both the provider and the claims file, and this is routinely done.

⁵ Roughly 50 percent of the Medi-Cal population is enrolled in a managed care health plan. The enrolled population consists primarily of beneficiaries constituting the family aid code categories. The aged, blind, and disabled population represents an optional category for most forms of managed care and is not enrolled in significant numbers. [See California DHCS, Medi-Cal Beneficiaries by Age/Demographics.](http://www.dhcs.ca.gov/dataandstats/statistics/Pages/Age_Demographics.aspx)

Researchers throughout the United States have used the Medi-Cal dataset to study hospitalizations for ambulatory care sensitive conditions⁶ across payer types and to assess public policy⁷ and various other health-related public policy issues.⁸

The Medi-Cal dataset can and has been linked to the California OSHPD hospital discharge data and to the vital statistics birth and death files. These linkages have been important for conducting AIDS research, assessing ambulatory care sensitive conditions admission rates, and evaluating birth outcomes.

Office of Statewide Health Planning and Development Hospital Discharge Data

OSHPD hospital discharge data is the only “comprehensive” health data file collected by California government that captures individually identifiable health information between birth and death. Each quarter, hospitals provide to OSHPD the hospital discharge dataset, which contains the discharge diagnosis and all billable and related procedures provided the individual during the hospital inpatient stay. “Comprehensive” is in quotes, however, because even though the dataset covers all inpatient hospitalizations in the state, not all California residents visit a hospital during their lives; for example, someone born at home who was never subsequently hospitalized would have no inpatient record. This dataset includes elements such as beneficiary city, state, county, ZIP code, gender, and ethnicity; principal language spoken; date and place of admission; and a Social Security Number that is converted to a unique identifier prior to internal or external use. Except for the Medi-Cal dataset, the state does not have a comprehensive dataset covering the full continuum of care, the setting for the bulk of medical care provided.

Environment

To fully understand and assess the barriers to accessing community-level data, one has to view the situation through the eyes of the governmental data steward. The complex political environment in which he or she operates is ever changing and filled with both actual and perceived personal and professional risk.

An understanding of this environment is also important when assessing the difficulties and roadblocks encountered when trying to access governmental data holdings and can help explain the resistance encountered when requesting such data. The governmental data steward—who has been entrusted with data sources that contain protected health information—must continually

⁶ Andrew B. Bindman, Arpita Chattopadhyay, Dennis H. Osmond, William Huen, and Peter Bacchetti (2005, February), “The Impact of Medicaid Managed Care on Hospitalizations for Ambulatory Care Sensitive Conditions,” *Health Services Research*, 40(1), 19–38.

⁷ [National Bureau of Economic Research \(2002, August\)](#), *Does Contracting Out Increase the Efficiency of Government Programs? Evidence from Medicaid HMOs* (Working Paper No. 9091), [Mark Duggan](#).

⁸ [Tania Barham](#) and [Paul Gertler \(2006, June 4\)](#), *Making Babies Healthier by Providing a Managed Care Option to California's Poor*, paper presented at the annual meeting of the Economics of Population Health, Inaugural Conference of the American Society of Health Economists, TBA, Madison, WI.

assess the political and legal environments, which are in a continual state of flux.⁹ The environment in which the governmental data steward must work creates a culture that is risk adverse. If the steward releases data that is subsequently disclosed to unauthorized individuals, he or she may undergo extreme scrutiny. The government data steward's agency may also become a target for political folly. Recent data breaches in California that led to the enactment of new laws and castigation by political representatives present a prime example of this environment. To further complicate the landscape, many of the government datasets contain health information associated with vulnerable populations that may be subject to discrimination or ostracized should their health data be accessed inappropriately. And many health care data breaches are splashed across newspapers, creating a sense of fear among the public. In fact, surveys have shown that roughly 70 percent of the public is concerned about the privacy of its medical records, and in some cases this is warranted.¹⁰

As a California state senator (D-Redondo Beach, 28th District), now secretary of state Debra Bowen introduced Senate Bill 13 (Chapter 241, Statutes of 2005) in response to a data breach at the University of California, Berkeley. In 2004, a hacker accessed the computer of a UC Berkeley researcher and information on more than 1.3 million persons who received or provided services to the state's In-home Supportive Services Program. According to Sen. Bowen, Senate Bill 13 was implemented to provide greater oversight and protection of individuals' confidential information.

When Sen. Bowen introduced the legislation, she stated:

Identity theft is still the country's fastest-growing white-collar crime, and it's maddening to see state agencies responsible for handling sensitive personal information still don't understand that a person's Social Security number is the one key criminals need to unlock someone's entire financial history. The state needs to take a hard look at its data-sharing laws to make sure Social Security numbers and other key data identity thieves thrive on aren't being handed out like holiday eggnog.¹¹

The follow-up investigation led to the conclusion that the data steward should not have released such a "large" dataset. Instead, the data steward should have worked with the requester to develop a sample that would have limited the department's risk. The Department of Social Services was required to notify the individuals affected and incurred costs totaling \$700,000. In addition to the legislative response, all departments were asked to examine their internal policies about data sharing. The California chief information officer, Clark Kelso, stated that "we shouldn't be sharing information unless there's a compelling research need for it."¹²

⁹See, for example, Appendix B, which displays the volume of California legislation associated with information protection and/or privacy for fiscal years 2005-06 and 2006-07.

¹⁰ For a discussion of medical identity theft and how it can harm its victim, see Pam Dixon (2006, Spring), "Medical Identify Theft: The Information Crime That Can Kill You," World Privacy Forum.

¹¹ Thomas Claburn (2004, December 6), "Proposed California Bill Bans Distribution of Social Security Numbers," *InformationWeek*, <http://www.informationweek.com/story/showarticle.jhtm?articleID=54800697>.

¹² Claburn, "Proposed California Bill."

In a case that drew national attention, veterans were exposed to potential identify theft when a data analyst took home a laptop containing 26.5 million veterans' personal information, and the laptop was stolen. The news accounts noted that the unidentified analyst was placed on leave pending review.¹³

The public's perception is also a key element. Surveys disclose that the public is concerned about the privacy of its personal medical records. Any data breaches, no matter how insignificant, will draw the ire of not only the public but also its political representatives. The California HealthCare Foundation (CHCF) commissioned a national consumer health privacy survey in 2005, which revealed that 67 percent of the national respondents were "somewhat" or "very concerned" about the privacy of their personal medical records. This concern was even greater among people of color: 73 percent were "somewhat" or "very concerned." The public's trust is an important factor for both the requesters of the data and those entrusted with its security, because without the public's trust, democratizing data and creating lasting data-sharing policy will be difficult to accomplish.

Data is used to make governmental decisions in this country every day. It's used to allocate money, divide up political districts, marshal resources to combat public health threats, and plan communities, and for a whole host of efforts designed to enhance our communal life. These uses generally go unnoticed by the populace and cause little concern. But data and research endeavors can also abuse and harm certain segments of our society.

The Tuskegee syphilis study is probably the most egregious example in the United States of this abuse. For 40 years, 399 African-American males were deceived by United States public health officials and denied treatment for syphilis in a study performed to document the natural history of disease. The participants were never told that they suffered from syphilis. Instead, they were told that they were being treated for "bad blood," and the physicians who studied them went to "extreme lengths to insure that they received no therapy from any source."^{14 15}

The outcome of this event was mistrust. The African-American community became wary of medical research and was less willing to share its health information. Some have suggested that this event even contributed to the shortage of organs used in transplantation for the African-American community.¹⁶ This legacy of mistrust has major implications for data sharing.

¹³ Martin H. Bosworth (2006, May 22), "VA Loses Data on 26 Million Veterans," ConsumerAffairs.com, http://www.consumeraffairs.com/news04/2006/05/va_laptop.html.

¹⁴ University of Virginia Health System, "Final Report of the Tuskegee Syphilis Study Legacy Committee—May 20, 1996," http://www.hsl.virginia.edu/historical/medical_history/bad_blood/report.cfm.

¹⁵ University of Virginia Health System, "Bad Blood: The Tuskegee Syphilis Study," http://www.hsl.virginia.edu/historical/medical_history/bad_blood/.

¹⁶ Stephen B. Thomas and Sandra Crouse Quinn (1991), "The Tuskegee Syphilis Study, 1932–1972: Implications for HIV Education and AIDS Risk Programs in the Black Community," *American Journal of Public Health*, 81, 1503.

Recent events, such as the cancellation of health care policies by Blue Shield of California, also help create a climate of mistrust and concern over how medical information is used. During an investigation, the California Department of Managed Health Care discovered that Blue Shield of California hired a team of analysts who mined the data to identify high-cost enrollees. Blue Shield allegedly combined the mined information with pre-enrollment screening data to identify individuals who did not disclose preexisting conditions¹⁷ and then dis-enrolled the individuals and refused to pay their medical expenses. While Blue Shield did not admit guilt, it did reinstate the 682 enrollees' policies and agreed to pay all medical claims for these individuals.¹⁸

The common thread in such recent horror stories is that an individual's health care data—either captured in a prescreening document or through the claims-processing system—was used to rescind health care coverage. Understandably, the idea of losing one's health care coverage while gravely ill greatly disturbs the public. Would you share your health information if your action could result in the recession of your or your child's coverage?

This complex environment plays a significant role in determining whether the data requested will be released or severely restricted in its content to limit risk to both the government organization and the data steward. As discussed above, data breaches tend to result in second-guessing. The data steward may have followed all of the policies, but the hysteria and political risk associated with a data breach result in follow-up investigations focusing not only on what went wrong but also on what should have been done. When this environment is considered in light of the fact that all such data releases are permissive, it becomes evident why data stewards are reluctant to take on the risk associated with the data release; it generally outweighs any benefits that may accrue.

Limited Policy Direction

Legislative bodies have responded to data breaches and abuses with legislation designed to prevent such occurrences. Institutional review boards have been created at various levels of government and at research institutions. Legislation specific to governmental departments, governmental programs, and even specific governmental datasets has been enacted, in most cases in reaction to specific events or practices or to protect vulnerable populations. At present, no central authority exists to govern or compile the policies governing data releases for each of the health care datasets discussed in this paper.

In many cases, data-release policies are not consistent across departments. Often both federal and state statutes govern the data, making any decisions regarding releases difficult. The most complicated situation occurs when two datasets are matched. The data steward then must apply the state requirements for each of the distinct datasets plus any applicable federal requirements when making decisions about releases. The result is a complex, confusing, and potentially hostile environment for the data steward.

¹⁷ Lisa Girion (2006, September 17), "Sick but Insured? Think Again," *The Los Angeles Times*.

¹⁸ Paul Krugman (2006, September 22), "Insurance Horror Stories," *The New York Times*.

In this section, we highlight, from a policy point of view, some of the difficulties data stewards encounter when trying to release combined datasets. As an example, we use an actual data request that was submitted to the DHCS and the OSHPD.

OSHPD and DHCS Combined Data Request Case

A university researcher requested hospital patient discharge data from OSHPD that he intended to match with Medi-Cal eligibility data. The researcher proposed to study hospitalizations for ambulatory care sensitive conditions by payer and geography. He further intended to utilize geographic information system (GIS) software to prepare maps and display areas that may be lacking adequate access to primary care, thus requiring the patient address to be included in the dataset.

We found that the first complications encountered were the differing governing statutes for releasing each dataset. First and foremost among these was the Social Security Act Title XIX restriction, which limits Medicaid data releases for “purposes directly connected to the administration of the program.” Also pursuant to Title 42 Code of Federal Regulations Section 431.301, Medi-Cal data can be released only for purposes of these direct administrative concerns, which include:

- Establishing eligibility
- Determining the amount of medical assistance
- Providing services for recipients
- Conducting or assisting an investigation, prosecution, or civil or criminal proceeding related to the administration of the plan

Although the statute does provide some leeway on this, broadening the interpretation increases the risk that the department and the data steward assume. Therefore, the current practice is to adhere closely to the enumerated permissions. As one can see, these restrictions pose many problems. For example, in the present case, the researcher first approached OSHPD requesting the information and was informed that OSHPD did not have access to Medi-Cal data. The researcher probably could have received the OSHPD hospital discharge dataset, assuming he or she complied with the Information Practices Act,¹⁹ since it does not require the release to be directly connected to the administration of any program. Initially the researcher had not contemplated this requirement; therefore, he was unable to access the Medi-Cal data. But once he had established that his research was directly connected to the administration of the Medi-Cal program, he embarked on yet another odyssey, because combining the datasets now became an issue.

To accomplish the research, a common identifier was required in order to match the two datasets. The primary variable available in the two databases was the Social Security Number, augmented by other variables, such as sex and date of birth. But most governmental organizations are extremely cautious about releasing Social Security Numbers, as a result of the increased

¹⁹ In addition to OSHPD policies and procedures governing data releases, the hospital discharge data release would be governed by California Civil Code Sections 1798.24 through 1798.24b.

occurrence of identify theft and the laws that have been adopted in response to the crime.²⁰ California Senate Bill 13 (Chapter 241, Statutes of 2005) incorporated provisions into the California Information Practices Act, in Section 1798.24(t)(2)(E), suggesting that government entities should attempt to minimize risk when possible. Specifically, this section states: “If feasible, and if cost, time, and technical expertise permit, require the agency to conduct a portion of the data processing for the researcher to minimize the release of personal information.” At this point in the case at hand, the data steward was faced with a decision: Should he or she release the Social Security Numbers and allow the researcher to perform the match? Or was it possible for the data steward to accomplish the match prior to release and then encrypt the Social Security Numbers? Again, the data steward had to consider several issues. If he or she decided to release the data with Social Security Numbers and allow the researcher to perform the match, and the data was subsequently breached, the data steward’s actions would certainly be second-guessed. A retrospective review would likely challenge the data steward’s conclusion that his or her organization did not have the time or expertise to conduct the match, concluding that the steward and thus the department erred. At this point, the data steward assessed risk, workload, and reward: Did the steward and the organization want to risk a data breach when they were not compelled to release the data? In the case we reviewed, the answer, not unexpectedly, was no. The steward decided not to release the data.

Our subsequent discussion disclosed even more issues that were not apparent to the researcher but were of significant concern to the data steward. Even if the steward had decided to perform the match for the researcher, he or she faced another significant set of issues, beginning with who specifically would perform the match. In this case, either DHCS or OSHPD could have performed it, but at some cost. Who would fund it? And once the match was performed and the dataset was turned over to the researcher, who would handle follow-up questions and coordinate the various departments’ responses?²¹

Finally, this data steward was aware, as were the others consulted during the case reviews, of policy inconsistencies among departments regarding the release of Social Security Numbers and did not wish to embark on such a quixotic mission. Because the two datasets were housed in two different departments, a real possibility existed that the other department might eventually decide not to release the Social Security Numbers, rendering the match impossible. While this was especially likely if the other department knew the steward would turn the data over to a researcher to perform the match, pitfalls existed even if the steward was planning to do the match in-house. In this case, the data steward receiving the data to perform the match was likely to be asked by the data-releasing department to complete a data use agreement prior to receiving the data for the match. While the steward had the authority within his department to sign off on releases of the department’s data, it was not clear that he or she had the authority to accept and hold confidential data from another department. In the bureaucracy, especially because of the

²⁰ California Senate Bill 13 (Chapter 241, Statutes of 2005) was implemented in response to a data breach that resulted in the unauthorized access to Social Security Numbers.

²¹ It should be noted that the release of a dataset generally results in a number of follow-up questions. Researchers and others working with the data inevitably run into data relationships that require input from those who collect it or those who are familiar with policy decisions that are reflected in the dataset.

data-release environment and workload priorities, these agreements may take weeks or months to execute, as there is generally no formal policy regarding how such a transaction should be handled. We found, in our case review, that it was not clear who would have been authorized to execute the agreement. Other issues, such as security protocols, were a concern as well. For example, if the department sending the data believes the receiving department does not meet the sending department's required security protocols, how should it proceed? In our case, the receiving department was reluctant to subject itself to scrutiny by the other department's staff.

Table 1 summarizes some of the major differences related to data releases associated with the three departments and the datasets studied. While not meant to provide the legal basis for data releases, it displays some of the difficulties that data stewards face when combining datasets. As can be seen, the datasets maintained by two of the three departments studied are not governed by HIPAA. The Medi-Cal data maintained by DHCS is the only dataset in our study that is governed by HIPAA. While HIPAA does apply to Medi-Cal data releases, the federal provisions governing data releases are more restrictive. The federal provision that data be released for purposes "directly connected to the administration of the program" does not apply to the other datasets. In addition, because Medi-Cal contains health data from several programs (e.g., California Children's Services, Foster Care Services, Alcohol and Drug Programs,²² Mental Health Services), which in many cases operate under statutes that specifically limit data sharing, additional policy issues enter into most data releases. For example, the Medi-Cal dataset includes claims associated with recipients of Alcohol and Drug Programs and from the California Children's Services program. Data releases from each of these programs must follow specifically adopted statutes governing the privacy of that unique dataset. Since the administration of these programs is housed either in a separate division within the Medi-Cal Program or in a different department, differing policies regarding data releases is likely. This means a researcher attempting to access the data must contact the data stewards of several departments and divisions to obtain approvals. Further, since each department's culture, view of data privacy, and even request and approval processes differ, time lengthens and the obstacles become insurmountable.

As noted in Table 1, the three departments studied each require a review and approval by legal counsel, an information security officer, a data steward, and, where applicable, a HIPAA compliance officer when confidential datasets are released. In many cases, the individuals responsible for these areas have differing interpretations of data security, data release policy, and other concerns specific to their individual domain. Much of this differing interpretation is driven by the organizational culture. Such things as recent data breaches, changes in executive management, and events that occurred to a particular person within the approval chain all influence data release policy. This differing culture inescapably leads to inefficiencies in data releases and results in additional costs for those attempting to access governmentally held data.

²² Title 42 CFR, Part 2—Confidentiality of Alcohol and Drug Abuse Patient Records sets forth specific confidentiality and data security standards for all medical records and data associated with all records maintained in connection with the performance of any program or activity relating to alcoholism or alcohol abuse education, training, treatment, rehabilitation, or research, which is conducted, regulated, or directly or indirectly assisted by any department or agency of the United States (Section 290dd-3 Confidentiality of patient records).

Table 1: Comparison of Data Release Policy Events for the Three Departments Studied

Policy	OSHPD Hospital Discharge Data	Vital Records Birth and Death Data	DHCS Medi-Cal Eligibility, Claims, and Provider Data
Is HIPAA applicable?	No	No	Yes
Does CA Information Practices Act apply?	Yes	Yes	Yes
Can data be released only for purposes directly connected to the administration of the program?	No	No	Yes
Is a minimum necessary?	Yes	Yes	Yes
Is CPHS approval required?	Yes	Yes	Yes
Does an internal committee review all data releases?	Yes	Yes	Yes, But relatively new
Is there a division or unit within the department with the express mission and resources for creating datasets to be released to researchers?	No	No	No
Does the department require its legal counsel to review each data release?	Yes	Yes	Yes
Must the information security officer review each data release?	Yes	Yes	Yes
Must the data	Yes	Yes	Yes

Policy	OSHPD Hospital Discharge Data	Vital Records Birth and Death Data	DHCS Medi-Cal Eligibility, Claims, and Provider Data
steward review the request and approve the data release as feasible and the identification of individuals?			
Must the HIPPA compliance officer review each data release?	No (Not applicable to this dataset)	No (Not applicable to this dataset)	Yes
Does it include data that are protected by statutes specific to a program, a disease, a condition, or other criteria (for example, HIV, Alcohol and Drug Programs)?	No	No	Yes
Do data release protocols inquire whether the researcher intends to display the data with GIS and how this will be accomplished to prevent original case locations from being disclosed?	No	No	No

Up to this point, our discussion of the data release case has focused on data matching but has not addressed the even more complex mapping component of the research endeavor. Our review found that the researcher who intended to use GIS software to analyze and display the ambulatory care sensitive condition information would have faced even more hurdles had his request progressed. Further, this situation was not unique to this specific request or department.

We found that none of the departments have established policy related to mapping of their data, although all were aware of and concerned about the issue. None of the departments were certain of the standards or the methodology they would employ to evaluate a request for geo-encoded data. And although the three departments expressed concern about this issue, none appear to reference this issue in their data request documents.²³ Further, they do not appear to have standards to employ when asked to approve data releases that might contemplate subsequent GIS mapping displays. Thus, since the stewards were aware of the potential for identifying individuals displayed on maps,^{24 25 26 27 28} each one expressed the opinion that he or she would “go slow” if required to contemplate releasing geo-encoded data at all. They were acutely aware that if a publication was released that allowed individuals to be identified, both the data requesters and the data releasers would be affected. If history provides an indication of the probable response, new laws would be likely, and department policies could easily be adopted that would further restrict a researcher’s ability to access data that can be mapped.

Lack of a Formal, Functioning Organizational Structure

An overarching problem that has begun to severely inhibit the availability of state health information began to emerge from our case studies. The problem stems from a lack of a formal body with the authority and staff to establish, maintain, and apply data policy standards. As we have shown with the three databases in our study, three separate organizations are responsible for both the data they collect and, more important, the policies and mechanics of releasing the data. Thus, there are three separate organizational cultures, three separate approaches to engaging with community leaders and researchers, and three different processes that requesters must negotiate. There are three different offices of legal counsel advising departmental staff on the releases, three different information technology security officers (ITSOs), three different chief

²³ See, for example, California Department of Public Health (CDPH) Data Request Forms, at <http://www.cdph.ca.gov/data/dataresources/requests/Pages/default.aspx>, or Office of Statewide Health Planning and Development (OSHPD) data request forms, at <http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html>. Researchers requesting confidential Medi-Cal data can access application material at <http://www.dhcs.ca.gov/dataandstats/data/Pages/AccessingProtectedData.aspx>.

²⁴ John S. Brownstein, Christopher A. Cassa, Isaac S. Kohane, and Kenneth D. Mandl (2006, December), “An Unsupervised Classification Method for Inferring Original Case Locations From Low-Resolution Disease Maps,” *International Journal of Health Geographics*.

²⁵ J Wartell and JT McEwen (2001), “Privacy in the Information Age: A Guide for Sharing Crime Maps and Spatial Data,” U.S. Department of Justice, Office of Justice Programs, <http://www.ncjrs.org/pdffiles1/nij/188739.pdf> [DEAD LINK].

²⁶ JS Brownstein, CA Cassa, IS Kohane, and KD Mendl (2005), “Reverse Geocoding: Concerns About Patient Confidentiality in the display of Geospatial Health Data,” *American Medical Informatics Association* [PLEASE VERIFY ORG NAME] *Annual Symposium Proceedings*, 905.

²⁷ JS Brownstein, CA Cassa, and KD Mendl (2006), “No Place to Hide—Reverse Identification of Patients From Published Maps,” *New England Journal of Medicine*, 355(16), 1741–42.

²⁸ AJ Curtis, JW Mills, and M Leitner (2006), “Spatial Confidentiality and GIS: Reengineering Mortality Locations From Published Maps About Hurricane Katrina,” *International Journal of Health Geographics*, 5, 44.

information officers (CIOs). None of these groups were reported to have special expertise in GIS or reidentification, leaving these issues to the domain of the data steward.

Compounding the impact of this variety, there are three (or more) different sources of funding that finance the programs that collect the data. Each of these provides different levels of support for staff to conduct necessary data activities; this means the priorities accorded to researchers will vary considerably by department at any given time, depending upon budget constraints and the varying demand for data activities brought on the department by external events. Few if any staff in any of these departments are devoted solely to handling issues with researchers, and therefore, when the demand to serve internal customers exceeds capacity, researchers and community groups usually fall to a lower priority. Since no central organization has established, standardized, and coordinated data request processes or handled the difficult business of reviewing requests and applying established standards before approving a data release, no central expertise has developed to allow efficient, accurate, and consistent processing. Each department and, in most cases, each program is left to its own devices to become proficient in recognizing issues inherent in a release.

We consider this lack of a central structure to be one of the key reasons GIS has been so slow to attain the success in California health departments that it has achieved in other fields. As we discuss in other sections, policy related to health information containing geo-encoded data records is limited or nonexistent. While solutions to thorny problems of identification, reidentification, and privacy when using geo-coded data are beginning to emerge in the greater universe of health data organizations, understanding and obtaining these solutions and, more important, applying them to actual datasets and data releases takes time and money. For each program, or even for each department, the cost to acquire the expertise and for the computer hardware and software and its related maintenance and operation is prohibitive. Further, even if an individual program did bear the expense and adopted a set of standards and a system for releasing de-identified geo-encoded data, such standards may be questioned or not recognized as adequate by others.

We have already discussed the case where the requester sought data from more than one department and the significant and thorny issues that arose. It is also clear that the added issues embodied in using and publishing geo-coded data make data releases even more complex. It bears repeating here that the lack of an overarching coordinating structure to develop and publish consistent application, to review and release standards, and to assist with establishing timely responses is a significant part of the problem. The additional risks and burdens that fall on an individual program data steward are usually more than the steward or his or her limited staff can handle, especially when the request requires reaching into another department's domain.

The balkanization of structures militates against development of common data release policies, techniques, staffing, standards, and equipment needed to consistently apply the standards and of a common orientation and economic efficiencies and a consistent orientation and culture that could better protect state data against unintended release while still affording adequate access to properly vetted researchers. Without a significant change in the current structure, the growing pressure from the external environment requiring departments to better secure their data, coupled

with decreasing resources and increasing risks and costs of handling external data requests, is likely to result in even more severely constricted data releases in the future.

Limited Resource Allocation

Governmental data stewards are surrounded to the point of being enmeshed in an extremely complex arena when contemplating data requests. To perform competently, they need resources. To data stewards, this means sufficiently trained staff, clear written policy standards regarding data releases, and adequate time to review requests against these standards. If data containing private information is to be released, it also means sufficient time and resources to properly vet the researcher and his or her environment to ensure adequate security can be maintained, and adequate computer systems and trained staff to operate them to accurately extract and provide the required data. None of the data stewards felt their resources were adequate to meet these criteria.

All the stewards contacted by the authors expressed essentially the same complaints.

- The majority felt that providing data to researchers was not a key part of their duties. The statutes that govern data releases principally center on data protection as opposed to mandating releases. The authority to release data at all is most often very general and is permissive rather than providing a mandate. As more than one steward pointed out: “When I’m faced with severe budget constraints where I am not even meeting my mandated duties under the law, on what authority do I spend time working on ‘permissive’ projects?” Since most governmental health datasets are collected for program administrative purposes and not for research or for community health advocacy purposes, such a response was of concern, but not a surprise.
- Another common complaint, expressed in a variety of ways, had to do with the complexity of the legal and social environment concerning health data and the personal and professional risk taken when providing data to individuals external to the department. Data stewards felt it was difficult enough keeping the data properly protected from inadvertent disclosure within their system, but attempting to ensure proper protection beyond their own domain involved substantial risk. The risk becomes higher with each news article on unauthorized disclosure and especially with each new law passed to “secure” data.
- Stewards reported that just getting through the process of data releases took an “inordinate” amount of time. In most cases, the researcher must be vetted to make sure he or she qualifies, since in many cases releases are confined by law to “a qualified institution of higher learning.” The most common interpretation of this limited the release to a person who is authorized to represent the institution, meaning that students or visiting faculty are excluded. Often a department head is required to support the acquisition of the data and ensure its protection. Verifying this requirement alone was sometimes time-consuming; then the request itself is scrutinized to make sure that each data element requested is necessary to the analysis, in accordance with the “minimum necessary” aspect in laws governing release. However, researchers understandably differ regarding these interpretations, especially when data substitutions designed to lower risk are offered (for example, “birth month” instead of “birth day”), requiring negotiation. Further, review by the department’s CIO, ITSO, and legal office is generally required.

Finally, computer operations must be assigned to extract the specific data fields required for the study from the database and copy them to appropriate media for secure transmission to the researcher. In many cases, this involves writing computer programs to modify the data before it is provided to the researcher. In most cases, state law and procedures require that these datasets be encrypted and that separate instructions be transmitted to the researcher in a secure fashion to allow access, yet another step in the process. Each data extract must be reviewed to make sure it complies with the final authorization. All of the above has to be documented and accessible in case of inquiry. Finally, since the data are actually “loaned” for a defined period rather than “given,” each dataset has to be tracked over time, and when the authorized period (usually one year) has expired, the data must be collected, a bona fide receipt of destruction must be received, or a reauthorization must be requested and supported by the researcher, processed through the system, and approved and documented.

- Stewards reported “personal aspects” to these endeavors. First, as shown above, most data releases are “permissive.” To data stewards, that means that if things go sour, hindsight will show that the data did not have to be released, with an immediate implication that the steward exercised poor judgment. Second, a number of recent laws, in certain cases, authorize the application of direct legal action and civil penalties against the person who released the data. While in most cases a governmental employee can expect the department’s legal staff to defend him or her in adverse actions, this is not necessarily a given in data release cases. Since the primary function of the legal staff is to defend the department and not individual employees, some uncertainty exists regarding how much protection the data steward can expect as an individual. Finally, with this considerable risk comes virtually no “reward.” Few data stewards are professionally recognized for their ability to provide data to external researchers. Rather, they are usually recognized for services they perform related to the department’s central mission and to program operations.

With increasing frequency, researchers, local health departments, and community groups are asking to receive information from matched health datasets. For example, matching vital record birth data with OSHPD hospital discharge data yields much more powerful information about the medical conditions, history, and outcomes surrounding births. However, as we’ve indicated above, performing such matches consumes a considerable amount of staff and computer resources. Faced with large demand and limited resources, data stewards often try to have the researcher pay for such matches and have external parties perform them. However, this usually adds to the risk the steward assumes, as the data is shipped “offshore”; creates more work for the steward, to vet the person doing the match; and results in a dataset “owned” by the researcher, so it generally is not available to others. And because converting the matched dataset to a de-identified set takes even more resources, it is often not done, and thus it is not available to the public.

It should be noted that the lack of resources does not necessarily mean that new resources are needed. However, it does mean that if current resources are to be used to perform these tasks, other mandated tasks must be repealed.

Conclusion

Researchers seeking health-related data compiled by government organizations are generally in for an eye-opening experience. While the government collects and maintains an impressive array of health-related datasets, gaining access to them is not for the faint of heart. In this paper, the authors discussed some of the major obstacles to democratizing these datasets. The significant barriers are generally related to the environmental context in which the governmental data stewards must operate, limited and sometimes conflicting policy direction, a lack of a centralized organization with an express mission for handling data and coordinating data releases, and limited resource allocation.

As discussed earlier, the governmental data steward works in a complex political environment that can be supercharged when data breaches occur. Stewards sometimes face the wrath of politicians who are attempting to react to their constituents' fears, as newspapers are emblazoned with headlines of data breaches from time to time. Sometimes this public fear and action on the part of policymakers is warranted, in response to valid holes in our protective public policy. Under such circumstances, additional laws and government policy are in order. But this environment also has its drawbacks.

| The current environment creates a risk-averse culture. When weighed against the perceived risk, any benefits associated with a given data release hardly measure up, especially since they seldom accrue to the department or the data steward. Thus, governmental data stewards and the department's review apparatus are inclined to either deny the request or require the researcher to severely modify the request to remove all risk (e.g., a de-identified dataset). This is especially true because the department and the governmental data steward rarely have the resources or the support they need to properly vet the request. Since such datasets are permissive, the pressure on government to stall, severely modify, or deny the request trumps release. While the effects of the environment can never be eliminated, formalized policy, an organization with an express mission to democratize government health data, and sufficient additional resources would go a long way toward the ultimate goal of democratizing government health care data.

Below is a list of recommendations we believe would greatly advance the democratizing data concept. These recommendations are focused on those barriers that we found inhibit government departments, data stewards, and all those involved from realizing the significant power inherent in governmental health data.

Recommendation 1: Create an Organization With an Express Mission to Coordinate Data Releases Across Multiple Departments and Datasets

As discussed in this paper, more than one governmental department maintains health care datasets. No one organization or department is responsible for coordinating data requests. In many cases, researchers and others who attempt to access governmental health care data resources must interact with more than one department. As a result, these data requesters must navigate through multiple departments that have designed, developed, and adopted unique policies and procedures. Still further, each department has its own unique culture that influences policy and creates inconsistencies among them.

The establishment of an organization or a department with an express mission to coordinate data releases among departments and datasets would solve many of these issues. This is especially relevant for combined datasets. As the example discussed in the policy section of this paper disclosed, matching two datasets that are maintained by two distinct departments is problematic. Resource and policy obstacles are almost insurmountable, and even when they are worked out, extensive delays in data releases are encountered. As previously discussed, many of the governmental datasets are subject to unique privacy rules, and even within a specific dataset (e.g., Medi-Cal), certain subpopulations of data are subject to unique rules, policies, and standards. Therefore, the proposed organization would work to establish the rules and policy for each unique dataset.

This proposed organization would be charged with identifying and specifying detailed business rules for the adopted policies and standards. Its mission would encompass establishing agreements with the departments that house the governmental health care data resources, including standards for privacy and security and rules for user authorization and access policies. It would maintain and publish a listing of all significant governmental health care datasets and establish and maintain meta-data for each. Standardized application forms must be developed and a standardized application process established and maintained. Priorities for accepting requests need to be established and published and should reflect health priorities established in the state health plan.

The organization must also develop the expertise to provide information that policy developers need to promote necessary changes to existing legislation and to address areas needing new legislation. For these changes to be successful, significant pressure must be brought by those external to state government who can best articulate the benefits that such changes will bring to the people.

When formulating this recommendation, we considered the fact that many departments that compile and house health care data would not want to relinquish their holdings to one central department. Therefore, our recommendation is that the newly developed organization would represent a standards- and policy-setting body. The separate departments would continue to compile and maintain their unique health care datasets, but the newly formed organization would be responsible for data release forms, proposal reviews, and the decision to release. Researchers would request data from this organization, which would maintain the requisite knowledge to help the researcher navigate to the correct dataset, to follow a standard process, and to apply the correct data release policies.

Recommendation 2: Develop a Funding Source for the Establishment and Operation of This Organization

To the extent possible, the organization should be supported by departmental programs that benefit from the research conducted and by research organizations. This would tend to insulate the organization from the worst of the vagaries of the annual state budget process.

Recommendation 3: Establish the Capacity to Match Datasets From Different Departments

Since matching datasets almost always requires the presence of identifiers, this function should be the responsibility of state government in order to limit the risk of data breaches. Further, the organization must have the capacity to provide the results to properly authorized departments in a format in which the requesting department can use it.

Recommendation 4: Develop, Maintain, and Operate Automated Systems to De-identify Data

These systems should employ specialized mathematical techniques to identify risks of identification or of reidentification from reengineering. A system should be created and maintained for use by departments to convert geo-encoded data into a form that can be used by mapping software but has been sufficiently altered mathematically to reduce the risk of reidentification. As we mentioned above, these systems are beginning to be developed and applied to health datasets, including those with geo-encoded data. However, the very newness of them, and the need to measure them for fit with the complex content and large size of the datasets we examined, means that this is a large, time-consuming, and likely relatively expensive proposition, and one that the current data stewards are unlikely to take on in the near future.