Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. 2015092903. September 29, 2015.http://techscience.org/a/2015092903

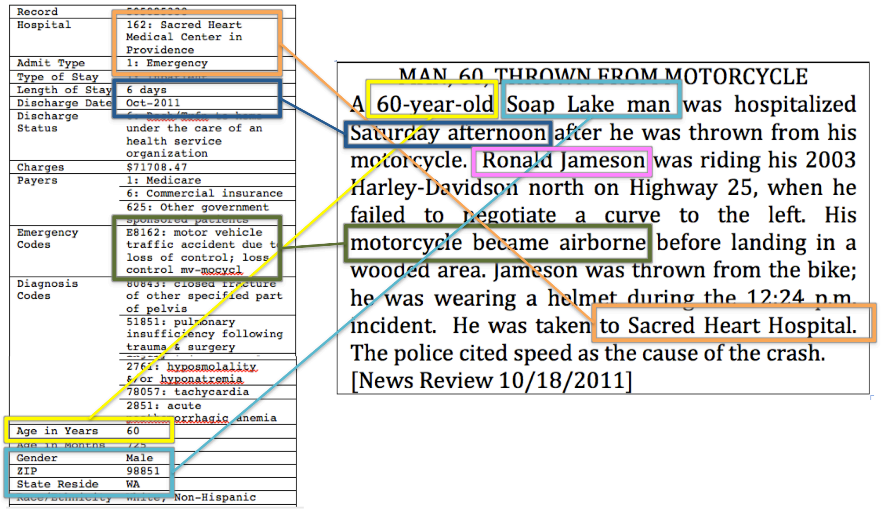
Technology Science

Only You, Your Doctor, and Many Others May Know

Latanya Sweeney

Highlights

- Washington State is one of 33 states that share or sell anonymized health records
- I conducted an example re-identification study by showing how newspaper stories about hospital visits in Washington State leads to identifying the matching health record 43% of the time
- This study resulted in Washington State increasing the anonymization protocols of the health records including limiting fields used for the re-identification study



Record	[illegible]
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	1: Inpatient
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	6: Dsch/Trfr to home under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare 6: Commercial insurance 625: Other government sponsored patients
Emergency Codes	E8162: motor vehicle traffic accident due to loss of control; loss control mv-mocycl
Diagnosis Codes	80843: closed fracture of other specified part of pelvis 51851: pulmonary insufficiency following trauma & surgery [illegible] 2761: hyposmolality &/or hyponatremia 78057: tachycardia 2851: acute posthemorrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]

Matching public medical information to news stories to identify patients.

Abstract

Alice goes to the hospital in the United States. Her doctor and health insurance company know the details — and often, so does her state government. Thirty-three of the states that know those details do not keep the information to themselves or limit their sharing to

researchers [1]. Instead, they give away or sell a version of this information, and often they're legally required to do so. The states turn to you as a computer scientist, IT specialist, policy expert, consultant, or privacy officer and ask, are the data anonymous? Can anyone be identified? Chances are you have no idea whether real-world risks exist. Here is how I matched patient names to publicly available health data sold by Washington State, and how the state responded. Doing this kind of experiment helps improve data-sharing practices, reduce privacy risks, and encourage the development of better technological solutions.

Results summary: The State of Washington sells a patient-level health dataset for $50. This publicly available dataset contained virtually all hospitalizations occurring in the state in a given year, including patient demographics, diagnoses, procedures, attending physician, hospital, a summary of charges, and how the bill was paid. It did not contain patient names or addresses (only five-digit ZIPs, which are U.S. postal codes). Newspaper stories printed in the state for the same year that contain the word "hospitalized" often included a patient's name and residential information and explained why the person was hospitalized, such as a vehicle accident or assault. A close analysis of four archival news sources focused on Washington State activities from a single searchable news repository studied uniquely and exactly matched medical records in the state database for 35 of the 81 news stories found in 2011 (or 43 percent), thereby putting names to patient records. An independent third party verified that all of the matches were correct. In response to the re-identification of patients in its data, Washington State changed its way of sharing these data to create three levels of access. Anyone can download tabular summaries. Anyone can pay $50 and complete a data-use agreement to receive a redacted version of the data. However, access to all the fields provided prior to this experiment are now limited to applicants who qualify through a review process.

Introduction

De-identification is the practice of removing name, address, and other explicitly identifying information from personal data. The idea of its protection is simple. If an individual cannot be identified in data, then the data can be shared freely without risk of harm to the individual. "Re-identification" occurs when you break de-identification by identifying an individual who is the subject of the data.

There are two irreconcilable belief systems posed by lawyers about current re-identification risks, and both are important to computer scientists, IT practitioners, and patients. Standard-bearer Paul Ohm, a professor of law at Georgetown University, asserts that useful data cannot be rendered anonymous in today's data-rich networked society [2]. If he is correct, a natural consequence is to stop trying technical fixes and seek non-technical alternatives. Opposing standard-bearer Jane Yakowitz of University of Arizona Law School asserts that no actual re-identifications have occurred and that prior claims have been overstated or misunderstood [3] [4]. If she is correct, a natural consequence is to accept current ad hoc de-identification as providing sufficient protection “as is” without a need for new tools or policy change.

Although they contradict each other, both standards discourage technological improvements. Differential privacy, which guarantees limited re-identification, has emerged as a dominant area of privacy research among computer scientists [6]. However, if differentially private tools were widely available today, neither standard would encourage their adoption. In order to improve the use of privacy-enhancing tools, we have to help society understand real-world risks and harms.

As a working example, I turn to publicly available data about personal hospital visits. Often medical information is benign — a broken arm gets a cast — but other times a hospitalization can include surprising results, such as drug or alcohol dependency appearing in an emergency hospitalization following a motor vehicle accident. Therefore, care must be taken when sharing patient information.

Years ago, most states passed legislation requiring hospitals to report to information about each patient’s hospital visit; most of those states, in turn, share a copy of the information widely for many purposes [1]. Actually, anyone can usually get a public version of the data that includes patient demographics, clinical diagnoses and procedures, a list of attending physicians, a breakdown of charges, and how the bill was paid for each hospitalization in the state. The information does not contain patient names but often includes ZIP codes, which are postal codes in the United States.

Statewide databases have been around for years and shared widely. If there was a problem, one might expect to be able to point to a litany of harms. As of this writing, I found no reported privacy violations from any state database, though it is unclear how and to whom one would report a violation. Most people are unaware of these statewide datasets, so even if harmed, it is unlikely an individual would be able to link harms back to the shared data.

On the other hand, there is anecdotal evidence. In a 1996 survey of Fortune 500 companies, one third of the 84 respondents said they used medical records about employees to make hiring, firing, and promotional decisions [7]. True or not, it is certainly possible, and the lack of transparency in data-sharing makes detection virtually impossible, even though the harm can be egregious. What is needed is a concrete example of how patients can be identified in this kind of data.

If you know someone who went to the hospital along with the approximate reason and/or the person’s general age, gender, and ZIP code, can you find his medical record in a state database?

At first glance, linking patients to publicly released health database records may seem academic or simply a matter of curiosity. But having an ability to access the records allows employers to potentially check on employees’ health, financial institutions to adjust credit-worthiness based on medical information, data-mining companies to construct personal medical dossiers, newspapers to uncover health information on public figures, and people to snoop on friends, family, and neighbors. Any of these parties could know when an individual may have gone to the hospital and other relevant information needed to locate that individual's health information in a public database.

States Are Not Covered by HIPAA

The Health Information Portability and Accountability Act (HIPAA) is a 1996 United States federal statute that authorizes sharing medical information, specifying to whom and how physicians, hospitals, and insurers may share a patient's medical information. State data collections are not subject to HIPAA because the states in their role as data collectors and disseminators are not entities covered by the HIPAA rule. Further, a state may share data mandated by state legislation in any form it deems appropriate. How do state decisions compare to HIPAA?

The Safe Harbor provision of the HIPAA Privacy Rule prescribes a way to share medical data publicly [8]. Dates may only include the year. HIPAA requires that ZIP codes contain only the first three digits if the population in those ZIP codes is greater than 20,000. ZIP codes for populations less than 20,000 report a null ZIP of 00000. No explicit identifiers such as name, Social Security numbers, or addresses can appear.

In comparison, only three states that share statewide hospital data do so in a manner that adheres to HIPAA guidelines; the other 30 do not [1]. Many states share health information with more specificity about admissions and discharges, such as providing the month and year of birth, instead of just the year. Other states generalize values beyond the HIPAA standard, such as providing age ranges and/or ranges of ZIP codes.

Background

Re-identifications of health data involve uniquely and specifically matching a named individual to a record, and re-identifications have been done previously. In 1997, I learned that medical information about state employees was slated for widespread sharing. The data holders removed explicit identifiers (e.g., name and address), in accordance with the best de-identification practice of the time. Other demographic information, such as date of birth, gender, and five-digit ZIP code remained. A mental calculation surprised me: There are 365 days in a year, two genders, and people live about 78 years. Multiplying these numbers gives 56,940 unique combinations. However, the average five-digit ZIP code in the United States has only about 25,000 people. To test my hypothesis, I needed to look up someone in the data. William Weld was the governor of Massachusetts at the time. His date of birth and home address in Cambridge were publicly known. For $20, I purchased the Cambridge voter list that included the name, address, date of birth, gender, and voting details for 54,805 registered voters [9]. Weld's demographics — birth date, gender, and ZIP code — were unique in the voter list and unique in the medical data, and combining them uniquely matched his identity to his record in the de-identified file released by the state about state employees.

News of the experiment spread to Washington, D.C., where policymakers were debating health privacy in what later became known as HIPAA. My first re-identification had a dramatic impact on the design of HIPAA's Privacy Rule, and I was personally mentioned in its preamble [10]. Discussion of my experiment also resulted in better protection for demographic values in regulations worldwide.

There have been several re-identification experiments since the Weld case. I had a litany of them immediately following the Weld experiment, but fear, shock, misunderstanding, and a lack of financial resources silenced results. For example, in *Southern Illinoisian v. Department of Public Health*, the Department confirmed that I had successfully re-identified children from {type of cancer, ZIP code, date of diagnosis}. While the court praised my skill and advocates dubbed me "the goddess of re-identification," the court ordered knowledge of my method sealed, barring me from publication. Similar fates awaited my other early re-identifications of survey and pharmaceutical data.

In the few experiments a decade ago where publication would have been possible, academic journals refused to do so for reasons having nothing to do with the scientific quality of the work. Computer-science publications refused to publish re-identification experiments unless the paper also included a technological solution, notwithstanding assertions that publishing these experiments would inspire technological innovation to address the real-world problem. Health-policy publications refused to publish re-identification experiments related to health data from fear that reaction might make data sharing more difficult, despite assertions that because technology was fostering unprecedented levels of data-sharing, it was timely to scientifically re-examine data-sharing practices. Even my Weld example and related demographic analyses, despite making significant contributions to privacy regulations worldwide, were refused publication by more than 20 academic publications at the time.

A decade ago, funding sources refused to fund re-identification experiments unless there was a promise that results would likely show that no risk existed or that all problems could be solved by some promising new theoretical technology under development. Financial resources were unavailable to support rigorous scientific studies otherwise.

Eventually, the void of published results fueled critics who wanted to assure the public that there were no such risks (even if, as was done in [3] [4], they created details that were distorted to the point of being inaccurate reports about the re-identifications).

Ten years later, El Emam et al. conducted a review and found only 14 published re-identification attacks [11]. Of the 14 examples, he and his co-authors dismiss 11 as being conducted by researchers solely to demonstrate or evaluate the existence of a risk, not to perform an actual verifiable re-identification. They classify the work of Narayanan and

Shmatikov [12] in this category. Narayanan and Shmatikov demonstrated the possibility of re-identifying published Netflix rental histories from the (identified) movie reviews submitted by Netflix customers. Regardless of the seeming dismissal of their experiment in the review by El Emam and his co-authors, Narayanan and Shmatikov's experiment led to an action by the U.S. Federal Trade Commission and a lawsuit that Netflix settled [5].

Of the remaining three actual re-identifications, El Emam and his co-authors dismiss two as having standards below those set by HIPAA. The authors promote the remaining study as being HIPAA-compliant and as having a very low risk of re-identification [13]; however, in the cited experiment, those researchers merely repeated the equivalent of my pre-HIPAA Weld experiment on post-HIPAA redacted data. The experiment over-fitted to what I had done and by doing so, failed to account for other possible re-identification strategies that may have been more successful. This is a critical shortcoming missed by El Emam and his co-authors. Nonetheless, the study by El Emam et al. underscores a need for better information on re-identification risks. This dismal state of scientific knowledge on re-identification risks beckons rigorous study in today's data-driven world even more than it did a decade ago. Therefore, can patients be re-identified in state health data today?

Methods

Materials for this experiment are: a collection of archival news stories; an online public records service for locating basic demographics on Americans; and a state database of hospitalizations in the same year and state as the archival news stories. More information about each resource appears below, followed by descriptions of preliminary processing of diagnoses and hospital codes and admissions dates.

News Stories

The LexisNexis newspaper archive [14] contains news stories printed in 2011 from a subset of Washington State newspapers. Searching the archive for stories containing the word "hospitalization" and that refer to a hospitalization of an individual yielded 66 distinct news stories from four sources: Spokesman Review (28 stories), The Associated Press and local wire (17 stories), The Columbian (19 stories), and The Mukilteo Beacon (two stories). Table 1 has a summarizes this information.

	Number of News Stories
Spokesman Review (Spokane, WA)	28
The Associated Press & Local Wire	17
Mukilteo Beacon	2
The Columbian (Vancouver, WA)	19
Total	66

Table 1. Distribution of news stories by news source for a total number of 66 stories.

Figure 1 provides a sample news story about a motorcycle crash that sent 60-year-old Ronald Jameson from Soap Lake, Washington, to Sacred Heart Hospital. (The person's name was changed in this example for privacy reasons.)

> MAN, 60, THROWN FROM MOTORCYCLE
>
> A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash.
>
> [News Review 10/18/2011]

Figure 1. Sample extract of a news story that contains name, age, residential information, hospital, incident date, and type of incident.

The news stories referenced 111 people. Some stories described incidents involving multiple people. Not all stories contained a name of an individual; only 86 names appeared in the news stories. One story did not have the names of the people, but did have a street address of a building where four people who lived there were hospitalized following a house fire. So, the total number of subjects is 90, which includes the 86 named people and the four people residing at the known street address.

Most news stories that listed names involved motor vehicle crashes (51 stories) and assaults (12 stories). Some stories reported medical hospitalizations (13 stories), primarily of well-known people or public figures (e.g., a professional soccer player, a judge, and a congressman). The remaining 14 stories reported shootings, suicide attempts, house fires, and other events. Table 2 lists the types of stories for the 90 subjects.

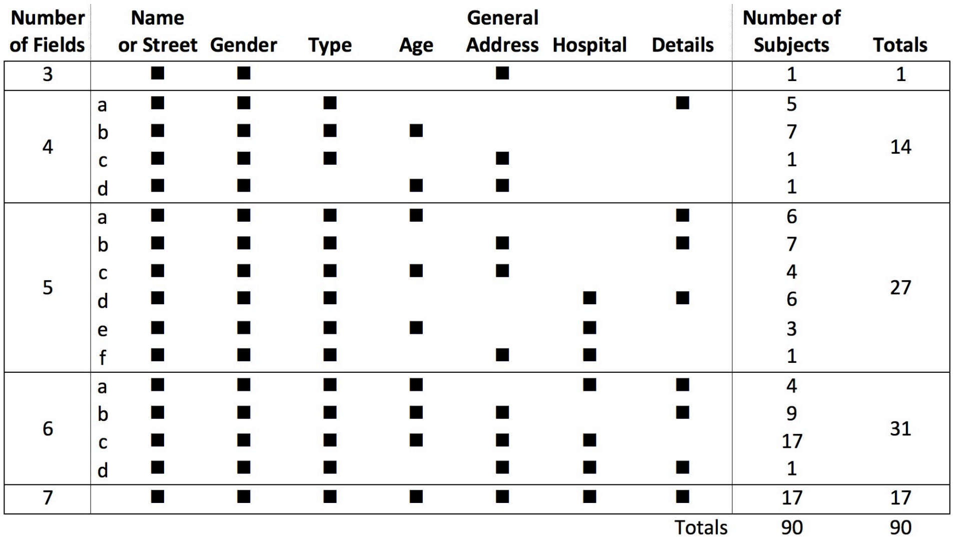
NEWS STORIES

	Number of Subjects	Percent
Motor Vehicle	51	57%
Assault	12	13%
Medical	13	14%
Other	14	16%
Totals	90	

Table 2. Distribution of news stories by type of incident for 90 subjects.

News stories tend to report the individual's name, age, residential information, type of incident, incident date, and hospital.

Harvesting these values, as available, from the news stories and adding the news source and publication date comprise the NewsData dataset used in this study. The dataset starts with 90 records, one for each subject. Figure 2 reports the distribution of fields in NewsData. Gender is present in all the records. Seventeen records have all the fields and 31 records have six of the fields. Only one record has just name, gender and address, with no hospital or medical content.



Number of Fields		Name or Street	Gender	Type	Age	General Address	Hospital	Details	Number of Subjects	Totals
3		■	■			■			1	1
4	a	■	■	■				■	5	14
	b	■	■	■	■				7	
	c	■	■	■		■			1	
	d	■	■		■	■			1	
5	a	■	■	■	■			■	6	27
	b	■	■	■		■		■	7	
	c	■	■	■	■	■			4	
	d	■	■	■			■	■	6	
	e	■	■	■	■		■		3	
	f	■	■	■		■	■		1	
6	a	■	■	■	■		■	■	4	31
	b	■	■	■	■	■		■	9	
	c	■	■	■	■	■	■		17	
	d	■	■	■		■	■	■	1	
7		■	■	■	■	■	■	■	17	17
								Totals	90	90

Figure 2. Distribution of values for fields harvested from news stories. Name is present in 86 cases, with four others having an explicit street address, for a total of 90 subjects. "General Address" refers to generalized residential information, such as town, city, county, or region of the state. "Type" and "Details" refer to the kind of incident and any medical details.

Online Public Records

Numerous online services offer search facilities for government-collected information (or public records) about a person in the United States. When an individual's name and/or other demographics are entered, these services may return the individual's date of birth, history of residential addresses, phone numbers, criminal history, and professional and business licenses, though specifics vary among states and services. Prices and results vary too. Some are free online, but most services offer a per-lookup fee (e.g., $3 to $10 per lookup). Some services offer monthly or yearly subscription plans, which can significantly lower per-lookup costs (e.g., 75 cents per lookup for up to 300 searches, or unlimited searches for $40/year). In this study, references to these kinds of search results are called PublicRecords.

Hospital Data

The Comprehensive Hospital Abstract Reporting System: Hospital Inpatient Dataset: Clinical Data [15] lists hospitalizations in Washington State for the year 2011 and cost $50. This data contains a record for each hospitalization in the state, and the total number of hospitalizations (or records) is 648,384. Each hospitalization record contains 88 fields of data. These include: the patient's five-digit ZIP code, age in years and months, race, ethnicity, and gender; hospital; month of discharge; number of days in the hospital; admission type, source, and weekend indicator; discharge status; how the bill was paid; diagnosis codes; procedure codes; and list of attending physicians. This dataset is termed HospitalData in this study.

For computer storage efficiency, most fields contain codes rather than English descriptions, so Washington State provides a dictionary defining each code.

Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	1: Inpatient
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	6: Dsch/Trfn to home under the care of a health service organization
Charges	$71,708.47
Payers	1: Medicare
	6: Commercial insurance
	625: Other government-sponsored payers
Emergency Codes	E8162: motor vehicle traffic accident due to loss of control; loss control mv-mocycl
Diagnosis Codes	80843: closed fracture of other specified part of pelvis
	51851: pulmonary insufficiency following trauma & surgery
	86500: injury to spleen without mention of open wound into cavity

	80705: closed fracture of rib(s); fracture five ribs-close
	5849: acute renal failure; unspecified
	8052: closed fracture of dorsal [thoracic] vertebra without mention of spinal cord injury
	2761: hyposmolality &/or hyponatremia
	78057: tachycardia
	2851: acute posthemorrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic
Procedure Codes	5781: Suture bladder laceration
	7939: 7919: Open/Closed reduction of fracture of other specified bone
Physicians	…
…	…

Figure 3. Sample extract of fields of information from a hospitalization record in HospitalData.

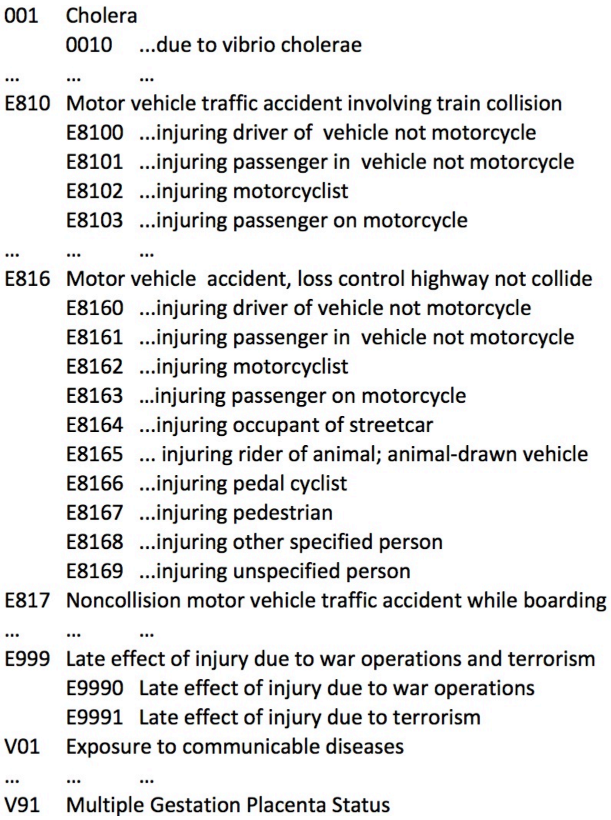
Figure 3 provides an example of part of a record in HospitalData, showing the code and its definition where appropriate. It describes an emergency admission (Admit Type field) of a 60-year-old, 725-month-old (Age fields), white, non-Hispanic (Race/Ethnicity) male (Gender). The total charge of $71,708 was paid by a combination of three entities. The emergency resulted from a motorcycle accident (Emergency Codes) and the diagnoses include a fracture of his pelvis (Diagnosis Codes).

Diagnosis Codes to NewsData

In a preliminary step, adding diagnosis fields to NewsData makes it more compatible for matching to HospitalData.

Using the type of incident described in a news story, we can list diagnosis codes that would likely appear among the diagnosis codes in HospitalData. The International Classification of Diseases, ninth edition (or ICD9), define more than 15,000 diagnosis codes [16] grouped into three categories: all numeric codes describe medical diseases (e.g., diseases of the digestive system or complications of pregnancy), codes beginning with an E describe external causes of injury or poisoning (e.g., motor vehicle accidents or assaults), and codes beginning with a V describe factors that may influence health status (e.g., communicable diseases, drug dependency, or tobacco use).

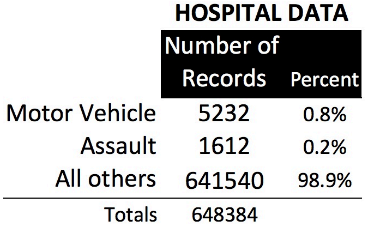
An ICD9 diagnosis code is an alphanumeric string where the leftmost characters represent a family of values, made more specific by adding more characters to the right. Figure 4 shows an example using emergency codes that begin with E81, which are an array of motor vehicle accident codes. A code of E816 describes an accident involving an out-of-control vehicle without a collision. Adding another digit provides more detail about how the patient was involved or injured. For example, E8162 describes a case where the vehicle was a motorcycle and the motorcyclist was injured.



001	Cholera	
	0010	...due to vibrio cholerae
...	...	...
E810	Motor vehicle traffic accident involving train collision	
	E8100	...injuring driver of vehicle not motorcycle
	E8101	...injuring passenger in vehicle not motorcycle
	E8102	...injuring motorcyclist
	E8103	...injuring passenger on motorcycle
...	...	...
E816	Motor vehicle accident, loss control highway not collide	
	E8160	...injuring driver of vehicle not motorcycle
	E8161	...injuring passenger in vehicle not motorcycle
	E8162	...injuring motorcyclist
	E8163	...injuring passenger on motorcycle
	E8164	...injuring occupant of streetcar
	E8165	... injuring rider of animal; animal-drawn vehicle
	E8166	...injuring pedal cyclist
	E8167	...injuring pedestrian
	E8168	...injuring other specified person
	E8169	...injuring unspecified person
E817	Noncollision motor vehicle traffic accident while boarding	
...	...	...
E999	Late effect of injury due to war operations and terrorism	
	E9990	Late effect of injury due to war operations
	E9991	Late effect of injury due to terrorism
V01	Exposure to communicable diseases	
...	...	...
V91	Multiple Gestation Placenta Status	

Figure 4. Excerpt of ICD9 Diagnosis coding tree. As more digits get appended to the right, the details get more specific.

More than half of the news stories involved motor vehicle accidents and a substantial number of stories described assaults (see Table 2). The ICD9 codes for motor vehicle accidents begin with E81 and E82. Assaults begin with E96. Table 3 reports the numbers of records containing these codes in HospitalData: 5,232 motor vehicle accidents and 1,612 assaults.



	HOSPITAL DATA	
	Number of Records	Percent
Motor Vehicle	5232	0.8%
Assault	1612	0.2%
All others	641540	98.9%
Totals	648384	

Table 3. Number of health records having a diagnosis starting with E81 or E82 for motor vehicle accidents and E96 for assaults from a total of 648,384 hospitalizations in HospitalData.

To facilitate simple matching of NewsData to HospitalData, I added diagnosis fields to NewsData and populated the fields with general versions of ICD9 codes (the leftmost digits) whose descriptions matched the incident details harvested from the news stories (see Details in Figure 2). In the 51 cases involving motor vehicle accidents, I merely recorded E81 and E82. In the 12 assault cases, I recorded E96. In the remaining 27 cases, I used an automated search for ICD9 codes whose descriptions matched details harvested from news stories. I recorded the most general version of the ICD9 code that matched the description — i.e., the three leftmost characters only.

For example, one news story reported Congressman Alcee Hastings was hospitalized with diverticulitis. A search of this term yielded ICD9 code 56211, so I added the code 562 to his record in NewsData.

As shown in Figure 2, all but two of the 90 records in NewsData had content in the Incident Type or Details field. I matched news descriptions to ICD9 codes in 72 of the 90 cases (or 80 percent of the stories) in NewsData. The specific diagnosis codes were: 437, 444, 508, 518, 562, 569, 800, 801, 802, 803, 804, 805, 808, 818, 824, 827, 829, 861, 864, 873, 884, 900, 910, 920, 923, 942, 943, 944, 945, 946, 947, 959, V58, E81, E82, E88, E89, E92, E95, E96, E97, and E98.

Hospital Codes to NewsData

Many of the news stories (50 of 90, or 56 percent, as listed in Figure 2) included the name of the hospital. One story merely referenced a Tri-Cities hospital, which is one in a group of about a dozen possible hospitals. As described earlier, HospitalData uses codes instead of English text in many fields, and one such field is hospital. An example appears in Figure 3. The code for Sacred Heart Medical Center in Providence is 162.

A dictionary of hospitals included with HospitalData lists 183 hospitals in Washington State along with their assigned codes. Some hospitals appear multiple times, with a letter added to the code to distinguish different units (e.g., rehabilitation or acute care).

An automated program compared the name of each hospital appearing in a news story to the dictionary of Washington State hospitals, added a new field to the dataset, and populated it with the correct code for the Hospital.

Nine of the hospitalization reports in the news stories were in other states, specifically Oregon and Idaho; as a result, those hospitalizations would not be in HospitalData. Therefore, these records were removed from NewsData, lowering the number of subjects to 81. The total number of subjects in the remainder of this writing will be 81 unless otherwise stated.

Other than those news stories referencing out-of-state hospitals and the one news story referencing a Tri-City hospital, all other hospitals in the news stories were uniquely matched and appended to NewsData.

Admission Dates to HospitalData

HospitalData includes the month and year of discharge and the length of stay in days but has no field for the admission date. On the other hand, a news story reports when an incident occurred. The date of the incident in the news story corresponds to the hospital admission date. So, I use the month of the discharge and the number of days in the hospital to compute a one-month range for the admission.

The earliest possible day of admission would be the first day of the month of the discharge less the number of days in the hospital (admitbegin). The latest possible day of the admission would be the last day of the month of the discharge less the number of days in the hospital (admitend). The date of the incident reported in the news story must be on or after admitbegin but before or on admitend for the news story to match that record of HospitalData.

For example, the news story in Figure 1 has an incident date of October 18, 2011. The medical record in Figure 3 reports a discharge month of October 2011 and a six-day stay. The earliest the admission could have occurred was September 25, 2011 (admitbegin), and the latest was October 25, 2011 (admitend). Therefore, the news story incident date matches the possible admission date for the medical record.

By now, I extended NewsData to include diagnosis codes and hospital codes based on information appearing in the news stories, and I reduced the number of records in NewsData from 90 to 81 by discarding news reports about out-of-state hospitalizations. I also extended HospitalData to include two fields, admitbegin and admitend, that describe a one-month window in which the admission must have occurred. Armed with these enhancements, I could match NewsData with HospitalData.

Approaches

As described earlier, a news story containing the word hospitalization often includes some combination of {name, age, residence, gender, hospital, incident} specifics sufficient to infer admission month and some of the diagnoses. An initial step is to look up the individual's name, age, and residence information in PublicRecords to learn the individual's date of birth and any five-digit *ZIP* codes associated with her. (This step can be automated with the purchase of a public records database.) Figure 5 provides a depiction of this initial step. Afterward, two different approaches were investigated, one involving automated matching of fields and the other using human exploration.

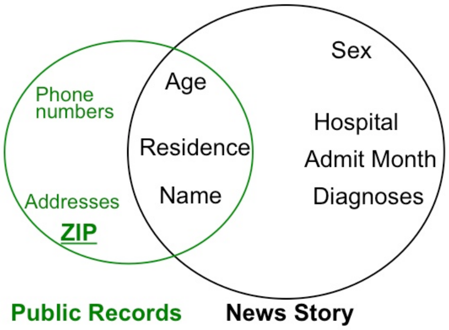


Figure 5. Acquire five-digit ZIP codes from public records using {name, residence information, age} from the news story. Age in years is from news and date of birth from public records.

Automated Approach

I wrote a computer program that makes a direct comparison between the newly learned *ZIP* and age in months (derived from birthdate and incident date) and other information from a news article — i.e., gender, age, hospital, admission month, and some likely diagnoses — to the fields of information in each record in HospitalData. My program did not use any blank fields in NewsData. When the overall comparison uniquely matches on each field of information provided, and the match of a news story is to one and only one record in HealthData, the result associates the name and residence information from the news story to the individual's health data, even though the health data did not previously include the name of the patient. If the comparison relates a news story to more than one record in HealthData, the comparison does not yield a match in this study. I accepted only exact and unique matches. If I found no match, the approach repeats, this time suppressing one or two values in the NewsData to see if one and only matching record appeared in HealthData. Figure 6 shows a depiction of the matching.

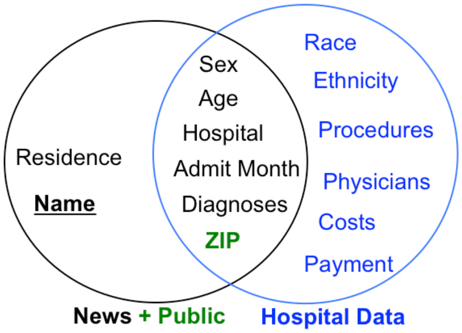
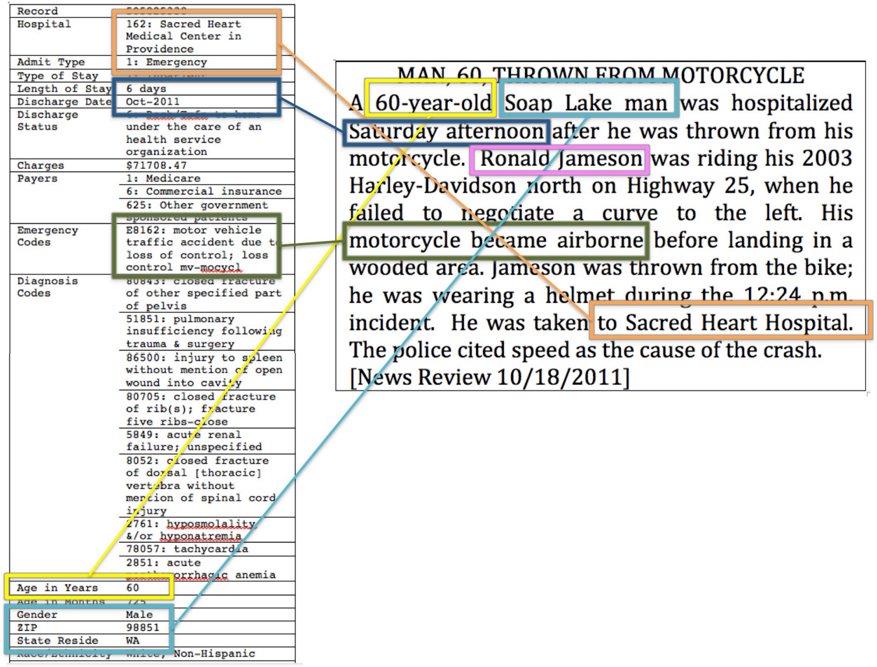


Figure 6. The automated approach matches news information and ZIP (from public records) to hospital data on a combination of {gender, age, hospital, admit month, diagnoses related to incidence, ZIP}, thereby putting a name to a medical record. Age is in years and months and the month of birth comes from public records.

For example, the news story in Figure 1 contains {Ronald Jameson, 60-year-old, male, from Soap Lake, admitted to Sacred Heart Hospital in October 2011 for a motorcycle-related accident}. If a search of HospitalData yields a unique match on these fields, then I consider the result a match of the news story to the hospital record. If there is no match, then I drop one field in the news story and retest, with the test repeating dropping each field in turn. If no records uniquely match when any one field is dropped, then I drop two fields and test each possibility. Figure 2 shows a uniquely matching health record. Figure 7 illustrates how the fields match.



Record	[illegible]
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	1: Inpatient
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	6: Dsch/Trfr to home under the care of an health service organization
Charges	$71708.47
Payers	1: Medicare 6: Commercial insurance 625: Other government sponsored patients
Emergency Codes	E8162: motor vehicle traffic accident due to loss of control; loss control mv-mocycl
Diagnosis Codes	80843: closed fracture of other specified part of pelvis 51851: pulmonary insufficiency following trauma & surgery 86500: injury to spleen without mention of open wound into cavity 80705: closed fracture of rib(s); fracture five ribs-close 5849: acute renal failure; unspecified 8052: closed fracture of dorsal [thoracic] vertebra without mention of spinal cord injury 2761: hyposmolality &/or hyponatremia 78057: tachycardia 2851: acute posthemorrhagic anemia
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash. [News Review 10/18/2011]

Figure 7. Example of information from news stories uniquely and exactly matching a medical record in publicly available Washington State health data.

Human Approach

A temporary employee, who was resourceful and knowledgeable about how to use the Internet and computers but who had no degree or training in computer science, mathematics, statistics, or medicine, was hired through an employment agency as a human investigator. I did not know her previously. After reviewing some of the matched and unmatched cases resulting from the automated approach, I asked the employee to see if she could find matching records for some of the new stories missed by the automated approach — i.e., cases where using all the fields in the news story led to no match. She could only use a set of candidate medical records, the news story, and any information she found publicly and freely available on the Internet.

Scoring

A news reporter agreed to score results by interviewing individuals who had been matched to the contents of the health record. (We did not publicly reveal any personal identity or medical information unless the individual explicitly agreed to share identity or medical information publicly.)

Results

Directly matching the fields of the records in NewsData to those in HealthData yielded unique and exact matches on 35 of the 81 qualifying records in NewsData (or 43 percent). Ten of the records in NewsData matched two records in HospitalData ambiguously, 11 matched three or more records in HealthData, and 25 matched none. We systematically reviewed the matches for consistency with details in the news story and other online information and no inconsistencies were found.

Of the 35 exact and unique matches, 30 matches used all the values supplied in the news story, five matches resulted when we dropped one value from the news story (ZIP, age, or hospital), and one match resulted when we dropped two values (age and hospital). In one case, we did not find the name that appeared in the news story, but a public records search without the last name matched one individual exactly and we used that name with her associated ZIP and age.

The scientific accuracy of these results rests on the accuracy of the newspaper and the health and public records information. Hospitals in the state must provide the information to the state by law, and the fields derive directly from billing records. HealthData reportedly contains all hospitalizations in the state [13]. Most of the news articles are from police reports; one was from a press release of the hospitalization of a congressman, and a few seemed to be regular news stories. There is no guarantee that these data are error-free, but any errors made in matching must result from errors in the data sources and not in the matching.

The automated approach uses an exact match, not a probabilistic one. Probabilistic matching would likely increase the number of matches, but my interest was to get as accurate a match as the data allowed, not as many matches as likely.

The news reporter was given the 35 matched results, from which he chose 14 to pursue by phone. He contacted eight, and confirmed all eight were correctly matched (100 percent correct). He attempted to contact six others, but was not able to make phone contact with them during the six-day evaluation period. Details from his interviews with those who agreed to public disclosure appear in his associated news story [17].

The human investigator was given two cases of high-profile people, a soccer player and a congressman, one case for which the news story was unusual (a sky-diving accident), and two other cases. No match was found using automated comparison for these cases without dropping one or two values. The human investigator had two workdays in which to investigate these five cases.

She compiled portfolios on each case, documenting not only which record in the HealthData was correct, but also why the exact automated comparison failed to match. She was successful in all five cases.

In the two cases of high-profile people, the ZIP code was not of a personal residence but the soccer franchise in the case of the soccer player, and his campaign headquarters in case of the congressman. When these ZIP codes were used, the modified news information uniquely and exactly matched the same health records the human investigator found.

The news story of the sky-diving accident found in LexisNexis had few details. When the investigator looked online at other news stories, she found the patient’s age and the fact that he lived in another state. Because he had an out-of-state ZIP code, she quickly was able to identify his record. When we appended his personal information to the news record, the automated comparison uniquely matched the same hospital record.

Similarly, her success in the other two cases resulted from augmenting the original news story with additional or correcting information she found online.

Most of the health records released by the state seemed to just report the specifics one would expect to find related to the reported incident. But about one-third of the records that we automatically matched to news stories (10 of 35 records) included references to venereal diseases, drug dependency, alcohol use, tobacco use, and other diagnoses or payment issues that may be sensitive, even though most of these records were for motor vehicle accidents.

Discussion

This experiment demonstrates how medical information for a targeted individual can be obtained using automated or human means, and neither methods requires sophisticated expertise. As one would expect, automated matching gives more results faster than the human approach, but the human investigator used other sources specific to the record.

There are many more newspaper sources available in the state. This study used only those available through one service at my university library, but other news services offer other sources, and most newspapers have their own websites. Overall, this means the number of possible cases drawn from news could be substantially larger.

Even though this study used newspapers as source information about a patient's identity, an employer could use hospital-leave information to check on employees, a financial institution could use credit account information to assess the credit-worthiness of clients who report illness as a reason for delayed payments, and any interested person could find out about the hospitalization of a friend or family member.

This experiment is important because it demonstrates how health data, currently shared publicly and widely without the knowledge of most patients, could put the privacy of patients at risk.

What Can Be Done

The goal is not to stop data-sharing. On the contrary, sharing data about patient encounters offers many worthy benefits to society. These data may be particularly useful because they contain a complete set of hospital discharges within the state, thereby allowing comparisons across regions and states of hospital and physician performance and assessing variations and trends in care, access, charges, and outcomes. Research studies that used these datasets include: examinations of utilization differences based on proximity [18], patient safety [19] [20], and procedures [21], and a comparison of motorcycle accident results in states with and without helmet laws [22]. The very completeness that helps these studies makes it impossible to rely on patient consent to sharing because the resulting data would not be as complete.

Another goal is to be smarter about how we perform data sharing. This is particularly important as the top buyers of statewide databases are not researchers but private companies, especially those constructing data profiles on individuals [23].

Washington State could choose to share its data in a form that adheres to the standards set by the HIPAA Safe Harbor, reporting dates in years and geography in three-digit ZIP codes (or blank if the ZIP has a small population). In patient-demographic fields, Washington's data has five-digit ZIP codes, the age of the patient, and a field that gives the age of the patient in months, which when reversed gives the year and a two-month window for the birth month. Washington's data also includes the month and year of discharge and the number of days hospitalized, allowing inference of a month range for the admission date. These extra inferences, five-digit ZIP code, admission month and bimonthly birth year, are more specific than just year, and helped in matching.

To be equivalent to HIPAA, Washington State could drop the discharge month and report no more than the patient's year of birth. Of course, these redactions may cause the resulting data to be less useful for some purposes. For those uses, Washington State could impose additional requirements for a data recipient to meet in order to acquire the more sensitive data.

After becoming aware of the experimental results, some states immediately began addressing the problem by improving the protections of publicly available statewide databases [24] [25]. States continue to impose stringent requirements on data requesters who need more identifiable data than these public versions. In particular, Washington State moved to provide three versions of the data, which differ by the amount of detail appearing in the data. Anyone can download tabular summaries. Anyone can pay $50 and sign a data-use agreement to receive a redacted version of the data. However, Washington State now limits access to all the same fields provided before this experiment to applicants who qualify through a review process [26].

Other options promise to come from technology. The technology that brought us today's data-rich, networked society is the same technology that can provide the best privacy protection. To get there, however, we must align policy and technology incentives, and that too is where this experiment fits in.

Policy should adopt best practices, which improve over time as privacy technology and the science of data privacy advances. Society can learn from cycles of published re-identifications, because the knowledge of vulnerabilities will rapidly lead to improved techno-policy protections. It is an evolutionary cycle. First, a re-identification vulnerability becomes known, which leads to improved practices and technical solutions, which in turn leads to other re-identifications, and so on, until eventually we achieve robust technical, policy, or administrative solutions.

The cyclic approach of expose-then-improve encouraged the development of strong encryption. Today, we use strong encryption for all sorts of tasks, such as online banking and purchasing. However, the earliest forms of encryption were just ad hoc decisions, similar to the kind of ad hoc decisions made about data-sharing today. In the past, someone would publish a way to break the leading scheme of the time, which spawned others to develop better methods, which in turn would be broken by others, until eventually we got the strong encryption society enjoys today.

Silence and fear break the development cycle in data privacy. Without an ability to learn about data sharing risks, knowledge stagnates and society blindly repeats the same errors in the face of increased technological vulnerabilities. One-day society might enjoy strong data privacy and the benefits of widespread data sharing but achieving this state requires rigorous open pursuit.

References

1. Hooley S, Sweeney L. Survey of Publicly-Available State Health Databases. Harvard University. thedatamap.org/1075-1.pdf

2. Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review, 2010. 57, 1701-1777.

3. Yakowitz J. Tragedy of the Data Commons. Harvard Journal of Law & Technology, 2011. 25(1), 1-66.

4. Barth-Jones D. The debate over re-identification of health information: what do we risk. August 20, 2012. [Web log comment]. Retrieved from http://healthaffairs.org/blog/2012/8/10/the-debate-over-re-identification-of -health-information-what-do-we-risk.

5. Netflix Prize Update. Netflix. March 12, 2010. http://blog.netflix.com/2010/03/this-is-neil-hunt-chief-product-officer.html

6. Dwork C. Differential privacy: A survey of results. In Theory and Applications of Models of Computation, TAMC 2008, volume 4978, pages 1–19. Springer, 2008.

7. Linowes D. A Research Survey of Privacy in the Workplace. White paper available from the University of Illinois at Urbana-Champaign. 1996.

8. Health Information Portability and Accountability Act (HIPAA) Safe Harbor Provision. 45 CFR 164.514(b)(2) (2002).

9. Cambridge Voters List Database. City of Cambridge, Massachusetts. Cambridge: February 1997.

10. The Privacy Rule of the Health Information Portability and Accountability Act. 45 CFR Part 160 and Subparts A and E of Part 164. www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/prdecember2000all8parts.pdf

11. El Emam K, Jonker E, Arbuckle L, Malin B. A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE, vol. 6, no. 12, Dec 2011, pp. 1-12.

12. Narayanan A, Shmatikov V. Robust De-anonymization of Large Spare Datasets. Proceedings of the IEEE Symposium on Security and Privacy, 2008, pp. 111-125.

13. Kwok P, Davern M, Hair E, Lafky D. Harder Than You Think: A Case Study of Re-identification Risk of HIPAA-compliant Records. Chicago, NORC 2011.

14. LexisNexis. www.lexisnexis.com/en-us/ product-finder.page
15. Comprehensive Hospital Abstract Reporting System: Hospital Inpatient Dataset: Clinical Data. Washington State Department of Health. http://www.doh.wa.gov/
16. ICD 9 CM : International Classification of Diseases: 9th: Clinical Modification 5th Edition. www.cdc.gov/nchs/icd.htm (For an online descriptive version, see www.icd9data.com/)
17. Robertson J. States Hospital Data for Sale Puts Privacy in Jeopardy. Bloomberg News. Also Bloomberg BusinessWeek. June 5, 2013. http://www.bloomberg.com/news/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy.html
18. Basu J, Friedman B. A Re-examination of Distance as a Proxy for Severity of Illness and the Implications for Differences in Utilization by Race/Ethnicity. Health Economics 2007;16(7):687-701.
19. Li P, Schneider J, Ward M. Effect of Critical Access Hospital Conversion on Patient Safety. Health Services Research 2007;42(6 Pt 1):2089-2108.
20. Smith R, Cheung R, Owens P, Wilson R, Simpson L. Medicaid Markets and Pediatric Patient Safety in Hospitals. Health Services Research 2007;42(5):1981-1998.
21. Misra A. Impact of the HealthChoice Program on Cesarean Section and Vaginal Birth after C-Section Deliveries: A Retrospective Analysis. Maternal and Child Health Journal 2007;12(2):266-74.
22. Coben J, Steiner C, Miller T. Characteristics of Motorcycle-Related Hospitalizations: Comparing States with Different Helmet Laws. Accident Analysis and Prevention 2007;39(1):190-196.
23. Robertson J. Who’s Buying Your Medical Records. Bloomberg News. June 5, 2013. www.bloomberg.com/infographics/2013-06-05/whos-buying-your-medical-records.html
24. Engrossed Substitute Senate Bill 6265. State of Washington. 63rd Legislature. 2014 Regular Session. http://apps.leg.wa.gov/documents/billdocs/2013-14/Pdf/Bills/Senate%20Passed%20Legislature/6265-S.PL.pdf
25. Patient Discharge Data. Healthcare Information Division. Office of Statewide Health Planning and Development. State of California. http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html
26. Comprehensive Hospital Abstract Reporting System (CHARS). Washington State Health Department.

http://www.doh.wa.gov/DataandStatisticalReports/HealthcareinWashington/HospitalandPatientData/HospitalDischargeDataCHARS

Authors

Latanya Sweeney is Professor of Government and Technology in Residence at Harvard University, Director of the Data Privacy Lab at Harvard, Editor-in-Chief of Technology Science, and was formerly the Chief Technology Officer of the U.S. Federal Trade Commission. She earned her PhD in computer science from the Massachusetts Institute of Technology and her undergraduate degree from Harvard. More information about Dr. Sweeney is available at her website at latanyasweeney.org.

The author gratefully thanks Lauren Mimran Ugbabe for her effort as a human investigator, Sean Hooley and Ryan Joyce for carefully and methodically reviewing materials, Jordan Robertson and Bloomberg News for validating results.

Editor: Pam Dixon

Citation

Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. 2015092903. September 29, 2015.http://techscience.org/a/2015092903

Data

Under review for data sharing classification. Data release available October 19.