Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901



Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data

Ji Su Yoo, Alexandra Thaler, Latanya Sweeney, and Jinyan Zang

Highlights

- We used newspaper data to match names to anonymized patient records in statewide hospital data from Maine and Vermont
- We found that 28.3 percent of names from Maine news stories and 34 percent of names from Vermont news stories uniquely matched to one hospitalization in the Maine and Vermont hospital data.
- When redacted to the HIPAA Safe Harbor standard, the Maine data allowed for a 3.2 percent re-identification rate and Vermont data allowed for a 10.6 percent re-identification rate.
- Our results suggest that states should revisit de-identification practices and reassess risks to patient privacy when determining data sharing protocol

Maine	NewsData	HospitalData
Demographics	Name	--
	19-year-old	Age: 19
	Female	Gender: F
	Bangor, Maine	Geocode: 19020
Hospital Information	York Hospital	HP: 200020
E-code	"hit by a car while crossing highway"	E8187 Other noncollision motor vehicle traffic accident injuring pedestrian
Diagnoses	"two broken bones in right leg, bruises to arms, legs, suffered head injuries"	87344 Open wound of jaw, without mention of complication
		82380 Closed fracture of unspecified part of tibia alone
		8249 Unspecified fracture of ankle, open
		88101 Open wound of elbow, without mention of complication
Sensitive information unrelated to hospitalization included in HospitalData		311 Depressive disorder
		30981 Posttraumatic stress disorder
		30000 Anxiety state

Re-identification example using details from news articles and de-identified hospital records from Maine

Abstract

Forty-eight states in the United States collect statewide inpatient discharge data that include personal health information of each patient's hospital visit [1]. A 2013 survey found that 33 of those states subsequently sold or otherwise disclosed copies of the data, but only three states de-identified data consistent with the standards established under the Health Information Portability and Accountability Act (HIPAA), the U.S. federal regulation that dictates the rules by which personal health information is shared [2]. While states are not mandated to follow HIPAA when de-identifying their data, many states still use a version of HIPAA to de-identify discharge data. Did the other 30 states put the privacy of personal health data at risk? To answer this question, Latanya Sweeney tested whether Washington State's hospital data was vulnerable to re-identification. The study showed that Washington State's inpatient data allowed for the correct matching of 35 of 81 (or 43 percent) individuals identified in the news stories to the anonymized discharge data released by the states local newspaper stories to anonymized hospital visits [3]. After the study, Washington State improved its anonymization standard for publicly available data and added an application process for others to receive more detailed nonpublic discharge data. Despite this successful outcome, many states were not convinced that the same re-identification strategy would be successful on their datasets. One reason was a belief that Washington State was more vulnerable because it shared patient age in months, a practice not followed by many other states. Is this correct? Are other states exempt from this re-identification strategy? To find out, we repeated the approach on statewide health data from 2010 in Maine and 2011 in Vermont using a total of 291 local news stories.

Results summary: We found that 69 of 244 (or 28.3 percent) of the names from local news stories uniquely matched to one hospitalization in the Maine hospital data. Even if redacted to the HIPAA Safe Harbor standard, the Maine data still allowed eight matches (or 3.2 percent). In Vermont, we found that 16 of the 47 (or 34 percent) names from the news stories uniquely matched to one hospitalization. If the Vermont data complied with the HIPAA Safe Harbor standard, five matches would result (or 10.6 percent). Such findings suggest that patients' personal information is vulnerable to re-identification even when hospital data is de-identified according to HIPAA Safe Harbor guidelines. We call for more rigorous inquiry on the vulnerabilities that exist even when following HIPAA Safe Harbor as a standard for de-identification.

Introduction

Forty-eight states collect and disseminate patient-level data for inpatient, outpatient, and emergency room visits that they then share with both private and public organizations [1]. Statewide collections of hospital data are useful to researchers, regulators, and businesses. Sharing a complete set of hospital visits within a state enables all kinds of analyses of hospital and physician performances [4]. Prior research studies have examined the effect of racial and geographical disparities on clinical trial participation [5]; studied the impact of critical access hospital conversion on patient safety [6]; analyzed the relationship between private health insurance and the increase in certain medical procedures [7]; and compared motorcycle accident results in states with and without helmet laws [8]. These are all worthy purposes to publicly provide patient-level statewide health data.

Before sharing the statewide collections of health data, each state decides how to de-identify patient-level data to protect privacy. These data include the patient's demographics, diagnoses, medical procedures, charges, and sources of payments for each hospitalization. Most states follow a version of the Safe Harbor method of de-identification found in the Health Information Portability and Accountability Act (HIPAA) Privacy Rule. HIPAA is the U.S. federal regulation that dictates the rules by which personal health information is shared by doctors, hospitals, and insurance companies. Statewide collections exist under the authority of state law and therefore, statewide hospital datasets are not subject to HIPAA. While states do not have to follow HIPAA, the Safe Harbor de-identification guidelines are still widely recognized as a standard for de-identifying all kinds of health data.

Under de-identification, a state removes all explicit identifiers, such as name and address, and replaces some specific values with generalized equivalents, such as reporting age instead of the day, month, and year of birth [9]. The resulting data has no name, address, or Social Security number and the only demographics that appear may be the patient's age, gender, and in some cases, a partial ZIP code. Some states, such as Maine and Vermont, apply their own alternative de-identification standard instead of adhering to HIPAA

requirements for de-identification. Maine and Vermont both share a patient's admission quarter, a date that is more specific than year.

To most people the resulting de-identified data can appear anonymous. Under this assumption, the data can be shared widely with researchers, private businesses, and the public without compromising personal privacy. But what if the state got de-identification wrong? What if even following the HIPAA Safe Harbor method for de-identification is not sufficient?

Assuming that states try to protect their residents' medical data privacy in how they de-identify their data, initial efforts at de-identification may not be successful, since risk-free de-identification is difficult. But diagnosis codes that are specific to one person's hospital visit and a patient's age, gender, and ZIP code may allow for the name of the patient to be associated with the de-identified record. The diagnosis codes that appear in the data describe the purpose of each visit and any health issues that may be relevant to or complicate care. For example, the record for a patient hospitalized with a broken arm because of a motorcycle crash could include a diagnosis code about his broken arm but also other codes describing alcohol, illegal drug use, diabetes, and other conditions. Therefore, even with the removal of explicit identifiers, patient records can be traced back to an individual.

In this paper we test two states' de-identification standards using publicly available statewide hospital data. Since Maine and Vermont have engineered their own alternative de-identification standards, we first test the level of risk in the state's standards. We also test HIPAA Safe Harbor standards using a version of the Maine and Vermont statewide datasets that adhere to HIPAA de-identification guidelines. If we demonstrate the risks to re-identification from HIPAA Safe Harbor or a state's current alternative standard, states have an opportunity to respond to improve their practices to reduce the risk in future data releases.

A cycle of research and improvement can improve states' de-identification standards and lower risk to patient privacy while still preserving the underlying utility of the data for public health and other purposes. However, in many states recipients of a copy of statewide dataset must sign an agreement that they will not re-identify patients. Some agreements include penalties that apply if the state learns that a recipient did re-identify patients. In practical terms, the data-use agreement gags the data recipient from talking about any re-identification vulnerabilities. It does not, however, necessarily stop re-identifications from occurring.

In fact, for those who may have the greatest incentive to exploit the data, a data-use agreement is not necessarily a deterrent. Among users of statewide hospital data are data analytic companies[10][11], many of which sell comprehensive, longitudinal datasets of

patient histories that include state hospital data. Financial incentives to link data could encourage re-identifications and outweigh any concerns raised by a data-use agreement. The data-use agreements can also discourage anyone from exposing the vulnerabilities since there is no requirement to report vulnerabilities. Data-use agreements that penalize re-identifications do not allow for data protection but re-identification studies such as this one can provide specific guidance on the existing weakness in de-identification methods and encourage states to implement stronger privacy protections.

Background

As mentioned earlier, 48 states collect patient information on each hospitalization that occurs in the state, and in 2013, 33 states subsequently disclosed copies of the data [1]. Of these states, only 3 states de-identified their data in a manner that was as rigorous as HIPAA requires [2]. The other 30 states shared data in a manner that was less protected than HIPAA warranted.

Washington State was one of these states. For $50, Sweeney purchased a copy of Washington State's publicly available dataset [3]. It had virtually all hospitalizations occurring in the state in 2011, and included 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 other than residential ZIP codes.

Newspaper stories in Washington State for the same year that contained the word "hospitalized" often included a patient's name and residential information and explained the reason for the hospitalization, such as vehicle accident or assault. This is the same kind of information an employer, family, friend or neighbor might know about a patient.

Sweeney assembled a sample of 81 news stories and found that news information uniquely and exactly matched medical records in the state database for 35 of the 81 sample cases (or 43 percent), thereby putting names to patient records. An independent news reporter verified matches by contacting patients [12]. Matches included high profile cases, such as politicians, professional athletes, and businesspeople. Some of the codes included sensitive information beyond the purpose of the visit, such as drug and alcohol use and sexually transmitted diseases.

After becoming aware of the experimental results, Washington State addressed the problem by requiring a data-use agreement of the publicly available version of the data and making a more detailed version available only through an application process [13]. California also changed its data disclosure practice [14]. Washington State found a means to provide detailed data to researchers and others while improving the privacy of the publicly available version.

In summary, a total of 30 states shared statewide hospital data in a manner less restricted than HIPAA's standard before the Washington State re-identifications. After the Washington State re-identifications became known, 28 states remained unchanged.

Two of these states are Maine and Vermont. Will the re-identification strategy used on Washington State hospitalization data work on Maine and Vermont's hospitalization data?

Re-identification

In this study, we use "named individual" to denote someone that we can identify by full name. In our re-identification strategy, someone who is identified by name in a local news article as being admitted to a Maine or Vermont hospital is considered a "named individual." We determine that there is a "unique re-identification" when a de-identified or anonymized record can be strongly associated with a named individual who most logically matches the details of the record." We use the term "binsize" to label the number of records that matches to a named individual. A "unique re-identification" has a binsize 1 (we write k = 1) where k is the number of de-identified hospital records that match to a named individual based on a combination of field values such as geocode, gender, age, etc. See Figure 1 for a visual depiction.

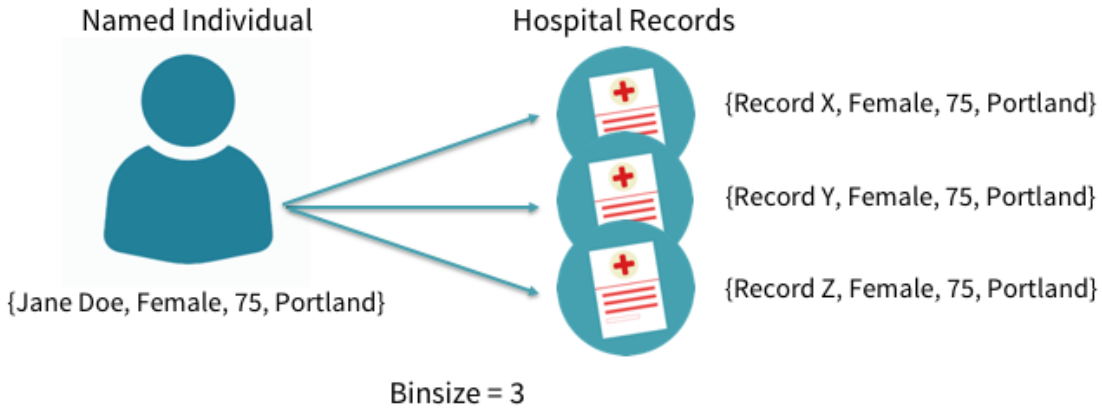


Figure 1. There are three hospital records that match the gender, age, and hometown profile of the named individual so binsize=3. If there was only one record of a 75-year-old female from Portland in the hospital records, then we call it a unique re-identification.

Hospital records often include the patient's previous and relevant conditions, even if those conditions have no relation to the reason for the current admission. If a male patient is hospitalized due to two-vehicle collision with a motorcycle, a re-identified hospital record may expose a diagnosis for a sexually transmitted disease. In cases where more than one

patient fits the details of a hospital record, all patients can still be strongly linked to that record.

Methods

The approach in the prior Washington State study [1] used a sample of local news stories that described accidents and incidents that led to hospitalization of named individuals because those kinds of news stories usually included the dates of visit, the patient's hometown, the name of the hospital, and the general reason for the visit — i.e., the same kind of information that others may know about a person's hospital visit The Bangor Daily News printed an example of such a news story: "James Smith, 64, of Bangor was treated at Eastern Maine Medical Center in Portland after a fire erupted when he lit a cigarette while using an oxygen tank." Smith was admitted to Eastern Maine Medical Center and his diagnosis codes likely related to burns and tobacco use.

In this study, we match news stories about named individuals to hospital visits to determine how successful we can uniquely associate a known patient to a de-identified hospital visit.

Materials

This study used a dataset of hospital visits occurring in a state within one year and a sample of news stories drawn from local newspapers from the same state and in the same year. The news stories describe named individuals hospitalized, primarily as a consequence of various accidents. The specific statewide datasets (MaineDataset and VermontDataset) and news stories used in this study (RawNewsStories) are detailed below.

MaineDataset

The Maine Health Data Organization (MHDO) has collected health care data since 1990 [15]. In 2013, we acquired a statewide hospital dataset called the 2010 Unrestricted Hospital Outpatient data from the MHDO for $1,125 [16]. The data purportedly contained all inpatient hospitalizations occurring in Maine hospitals in 2010 [17]. We refer to this dataset as the "MaineDataset."

Each row (or record) in the MaineDataset describes a single hospitalization; there are 105,808 rows. The fields of data associated with each hospital visit include: the quarter (January–March, April–June, July–September, or October–December) in which the hospitalization occurred; the patient's age, gender, race and ethnicity; coded values for the patient's hometown, hospital, diagnoses, medical procedures; the procedure dates (month, day, year) for medical procedures; status at discharge; point of admission; priority of the visit; type of insurance; and list of charges for the visit. There were 99 fields of data for each visit, all of which are listed in Appendix A. Figure 2 shows the 36 subset of fields used in this study.

Field #	Field Name	Field Description	Possible Values	Sample Record
1	mhdoid	MHDO Physical Record Number	QYYHPNNNNNNN	110240020795
2	hp	Hospital Code	6-digit code	200024
3	sx	Patient Sex	F,M	F
4	tadm	Priority of Visit (Type)	1-5,9	2
5	sadm	Point of Origin of Admission (Source)	1-9, B-F	4
6	adx	Admitting Diagnosis	ICD9-CM code	80131
9	geocode	Patient Geo Code	5-digit code	1050
10	disp	Patient Disposition/Discharge Status	2-digit code	62
17	dx1	Principal or First Diagnosis submitted	ICD9-CM code	80131
18	poa1	Principal Diagnosis Present on Admission	Y,N	Y
19	dx2	Second Diagnosis submitted	ICD9-CM code	E8809
21	dx3	Third Diagnosis submitted	ICD9-CM code	E8490
23	dx4	Fourth Diagnosis submitted	ICD9-CM code	5990
25	dx5	Fifth Diagnosis submitted	ICD9-CM code	7804
27	dx6	Sixth Diagnosis submitted	ICD9-CM code	2768
29	dx7	Seventh Diagnosis submitted	ICD9-CM code	4019
31	dx8	Eighth Diagnosis submitted	ICD9-CM code	30000
33	dx9	Ninth Diagnosis submitted	ICD9-CM code	3899
35	dx10	Tenth Diagnosis submitted	ICD9-CM code	
37	dx11	Eleventh Diagnosis submitted	ICD9-CM code	
39	ps1	Principal Procedure Code	ICD9-CM code	
40-44	ps2-ps6	Second-Sixth Procedure Code	ICD9-CM code	
45	pd1	Principal Procedure Code Date	YYYYMMDD	
46-50	pd2-pd6	Second-Sixth Procedure Code Date	YYYYMMDD	
61	los	Length of Stay		3

Field #	Field Name	Field Description	Possible Values	Sample Record
62	age	Patients Age		77
64	race	Patient's Race (coded)	1-5	5
65	ethn	Patients Ethnicity (coded)	1,2,8	2

Figure 2. List of 36 relevant data fields contained in the Maine statewide dataset for 2010 MaineDataset for each hospital visit occurring in the state. All reported diagnoses (i.e. fields "dx1," "dx2," ... "dx11") and procedures ("ps1," "ps2," ... "ps6") are alphanumeric codes found in the International Classification of Diseases, ninth revision (ICD9-CM). ICD9-CM codes indicate specific diseases, conditions, and procedures. The MHDO includes a data dictionary that defines demographic fields and hospitalization details that appear as numeric codes in the hospital record.

The MaineDataset has no explicit patient identifiers, such as name, address or Social Security number; it is de-identified. There are no explicit identifiers assigned to patients to indicate when multiple visits belong to the same patient. Each distinct patient visit is its own row in the data. Figure 3 shows an excerpt of one sample record from the MaineDataset with sample coded values explained using the MHDO dictionary [18] and the Research Data Assistance Center dictionary [19].

mhdoid	hp	sx	tadm	sadm	adx	geocode
110330020795	200033	M	2	4	8690	05170
Q = Jan-Mar, YY = 2010, HP= 33, NNNNNNN = unique number	Eastern Maine Medical Center	Male	Urgent	Transfer from another health care facility	Internal Injury To Unspecified Or Ill-Defined Organs Without Mention Of Open Wound Into Cavity	Portland

disp	dx1	poa1	ps1	los	age	race	ethn
1	E8132	Y	7817	4	60	5	2

disp	dx1	poa1	ps1	los	age	race	ethn
Discharge to Home or Self Care	motor vehicle traffic accident involving collision with other vehicle; mv-oth veh coll-motcycl	Yes	Application of external fixator device; tibia and fibula	In days		White	Non-Hispanic

Figure 3. Sample of one hospital record. Values in the sample are artificial.

The MHDO Physical Record Number Format ("mhdoid") is a unique number in the format "QYYHPNNNNNNN," where "Q"=admission quarter, "YY"=two-digit admission year, "HP"=two-digit hospital code, and "NNNNNNN"=unique integer assigned to each hospital record. The Hospital Code ("hp") is a unique, six-digit number assigned to each Maine hospital and clinic. The patient's gender ("sx") is recorded as "F" or "M" in the record, and race ("race") is coded in numeric values: "1"=American Indian or Alaska Native, "2"=Asian, "3"=Black or African American, "4"=Native Hawaiian or Other Pacific Islander, or "5"=White. Ethnicity ("ethn") has three possible values: "1"=Hispanic, "2"=Non-Hispanic, or "8"=Unknown.

The MaineDataset does not include patient identifiers such as name and address but it denotes a patient's "geocode." In Maine, a geocode is a five-digit number that includes a two-digit county number followed by a three-digit code that refers to the minor civil division [20]. For example, the county code for Kennebec County is 11 and the geocode for Augusta, a city in Kennebec County, is 11020. The Maine Office of GIS (MEGIS) provides a table of Maine geocodes that are used in the hospital data [21].

The MHDO lists 921 geocodes and 694 of them can be found in the MaineDataset. Figure 4 shows 20 geocodes with the highest number of records in the MaineDataset and the population of the geocode according to the 2010 U.S. Census [22].

Geocode	Town Name	Count in MaineDataset	Population in 2010
5170	Portland	8,470	66,194
1050	Lewiston	5,391	36,592
19020	Bangor	4,772	33,039
5220	South Portland	2,574	25,002
1010	Auburn	3,046	23,055

Geocode	Town Name	Count in MaineDataset	Population in 2010
31050	Biddeford	2,754	21,277
31230	Sanford	2,644	20,798
11020	Augusta	3,110	19,136
5200	Scarborough	1,663	18,919
31220	Saco	2,064	18,482
5240	Westbrook	2,221	17,494
5250	Windham	1,504	17,001
5090	Gorham	1,384	16,381
11240	Waterville	2,690	15,722
5030	Brunswick	2,399	15,175
31280	York	1,179	12,529
31110	Kennebunk	1,128	10,798
3530	Presque Isle	1,000	9,692
19050	Brewer	1,445	9,482
1060	Lisbon	987	9,009
31200	Old Orchard Beach	1,047	8,624
25280	Skowhegan	1,192	8,589
23020	Bath	1,235	8,514
3080	Caribou	1,179	8,189
7070	Farmington	1,008	7,760
9130	Ellsworth	1,251	7,741
13100	Rockland	1,222	7,297
27010	Belfast	977	6,668
3300	Houlton	1,053	6,123
23080	Topsham	1,049	5,931

Figure 4. Sample of 20 geocodes in Maine and the population per geocode. Highlighted geocodes have populations over 20,000. Towns are ordered by population.

2011 VermontDataset

We acquired the Vermont Emergency Department data from the HealthVermont.gov website. There was no cost for the data, and the website did not require any data-use agreement, credential, or application. We refer to this dataset as the "VermontDataset" [23]. Each record in the VermontDataset describes an individual emergency department encounter during 2011 in 14 hospitals. In 2011, there were 268,984 emergency department admissions in Vermont. We show the fields of data associated with each hospital visit below. There are 79 fields of data for each visit, which are all listed in Appendix B. Figure 5 shows the subset of fields we used in re-identifying Vermont's health data.

Field #	Field Name	Field Description	Possible Values	Sample Record
1	HNUM2	Hospital Number	1 or 2-digit code	5
2	ATYPE	Admission Type	0-5 (coded)	3
3	ASOUR	Admission Source	1 or 2-digit code	4
4	AgeGRP	Age Group	1-15 (coded)	14
5	ZIPGRP	ZIP code group	3- to 6- digit code	05478
6	SEX	Sex	0-2 (coded)	1
7	DSTAT	Discharge Status	0-13 (coded)	1050
10	DX1	Principal or First Diagnosis submitted	ICD9-CM code	80131
11	DX2	Second Diagnosis submitted	ICD9-CM code	Y
12	DX3	Third Diagnosis submitted	ICD9-CM code	E8809
13	DX4	Fourth Diagnosis submitted	ICD9-CM code	E8490
14	DX5	Fifth Diagnosis submitted	ICD9-CM code	5990
15	DX6	Sixth Diagnosis submitted	ICD9-CM code	7804
16-29	DX7-DX20	Second-Twentieth Diagnosis submitted	ICD9-CM code	2768
30	PX1	Principal Procedure Code	ICD9-CM code	4019
31-49	PX2-PX20	Second-Twentieth Procedure Code	ICD9-CM code	30000

Field #	Field Name	Field Description	Possible Values	Sample Record
50	ECODE1	External cause code	ICD9-CM code	3899
51	ECODE2	External cause code	ICD9-CM code	
52	ECODE3	External cause code	ICD9-CM code	
69	Uniq	Unique record number	8-digit	
70	ADMID_QTR	Admission quarter	1-4	
71	DISCD_QTR	Discharge quarter	1-4	

Figure 5. Sample list of the relevant fields in the Vermont statewide dataset for 2011. The dataset reports diagnoses (i.e. fields "DX1," "DX2," … "DX20") and procedures ("PX1," "PX2," … " PX20") as alphanumeric codes found in the International Classification of Diseases, ninth revision (ICD9-CM). HealthVermont.gov provides a data dictionary that explains the numeric codes. The Hospital Code ("HNUM2") is a unique eight-digit number assigned to each Vermont hospital and clinic. The patient's gender ("SEX") is recorded in coded numeric values: as "1"=Male, "2"= Female, or "0"=Unknown.

There are significant differences between the MaineDataset and the VermontDataset.

- The VermontDataset is twice as large as the MaineDataset because it contains emergency department encounters whereas the MaineDataset only contains inpatient admissions.

- In terms of data fields, VermontDataset reports age and location identifiers in more aggregated values than the MaineDataset. Maine provides Age in years while Vermont releases an aggregate "Age Group" for each record.

- Vermont also creates an aggregate "ZIP code group" category for patient location information. A patient with ZIP code group "052" could be from any ZIP code within the "05200-05299 range, excluding 05201." An excluded ZIP code is usually a large city in a ZIP code range. For example, 05201 is the town of Bennington with a population of 15,829. It is important to note that when a large city is excluded from the ZIP code group, both the remaining ZIP codes and the large city should then have a population over 20,000 in order to be listed as separate ZIP Codes in the data and still comply

with HIPAA Safe Harbor. This is not the case for all ZIP code groups in the VermontDataset as shown in Figure 6.

ZIP Code Group	Description	Count in VermontDataset	Population in 2010
37	ZIP in 03700-03799 range	1,178	77,452
50	ZIP in 05000-05099 range	9,764	61,355
51	ZIP in 05100-05199 range	14,321	29,595
52	ZIP in 05200-05299 range, excluding 05201	6,150	18,573
53	ZIP in 05300-05399 range, excluding 05301	6,378	20,458
54	ZIP in 05400-05499 range, excluding Burlington to Saint Albans	44,014	125,682
56	ZIP in 05600-05699 range, excluding Barre and Montpelier	22,486	56,237
57	ZIP in 05700-05799 range, excluding 05701	20,270	61,686
128	ZIP in 12800-12899 range	2,385	225,336
129	ZIP in 12900-12999 range	1,844	164,172
136	ZIP in 13600-13699 range	424	236,059
999	Non-USA or Unknown ZIP	734	
5201	Bennington	10,689	14,851
5301	Brattleboro	5,878	16,820
5401	Burlington	11,459	28,185
5403	South Burlington	4,868	17,593
5446	Colchester	4,361	15,373
5452	Essex Junction	4,119	19,710
5468	Milton	4,626	12,942
5478	Saint Albans	10,762	14,449
5602	Montpelier	4,771	11,916

ZIP Code Group	Description	Count in VermontDataset	Population in 2010
5641	Barre	10,021	17,169
5701	Rutland	11,899	20,802
5753	Middlebury	4,337	10,491
034+036	ZIP in 03400-03499 or 03600-03699 range	3,495	97,678
058-059	ZIP between 05800 and 05999	30,760	62,345
120-123	ZIP in 12000-12399 range	4,037	924,965
OTH-MA	All MA ZIP codes	2,765	
OTH-NH	All other NH ZIP codes	786	
OTH-NY	All other NY ZIP codes	1,942	
OTH-ST	All ZIP codes other than VT, NH, NY, and MA	7,461	

Figure 6. List of ZIP code groups in VermontDataset and 2010 Population in each ZIP code group according to the U.S. Census. Highlighted ZIP code groups have populations of more than 20,000.

Both the MaineDataset and the VermontDataset use standardized diagnoses and procedure codes from the International Classification of Diseases, ninth edition (or ICD9). The ICD9 defines more than 15,000 diagnosis and procedure codes [24] grouped into three categories: numeric codes that describe medical diseases (e.g., diseases of the digestive system or complications of pregnancy), codes beginning with an E that describe external causes of injury or poisoning (e.g., motor vehicle accidents or assaults), and codes beginning with a V that describe factors that may influence health status (e.g., communicable diseases, drug dependency, or tobacco use). Both the MaineDataset and the VermontDataset record each patient's diagnoses using ICD9-CM diagnosis codes. The MaineDataset includes procedure codes but the VermontDataset does not. We obtained a list of ICD9 diagnosis and procedure codes from the Center for Medical Service to interpret the conditions listed in the medical records. See Figure 7 for examples.

Type of Code	Code	Code Description
Diagnoses Codes	8208	Closed fracture of unspecified part of neck of femur
	9912	Frostbite of foot

Type of Code	Code	Code Description
	9490	Burn of unspecified site, unspecified degree
	29633	Major depressive affective disorder, recurrent episode, severe, without mention of psychotic behavior
	2910	Alcohol withdrawal delirium
	8690	Internal injury to unspecified or ill-defined organs without mention of open wound into cavity
Procedure Codes	8154	Total knee replacement
	7936	Open reduction of fracture with internal fixation, tibia and fibula
	9917	Injection of insulin
	9383	Occupational therapy
	8622	Excisional debridement of wound, infection, or burn
	9465	Drug detoxification
E-Codes	E8130	Motor vehicle traffic accident involving collision with other vehicle injuring driver of motor vehicle other than motorcycle
	E956	Suicide and self-inflicted injury by cutting and piercing instrument
	E9170	Striking against or struck accidentally by objects or persons in sports
	E0162	Activities involving building and construction
	E9220	Accident caused by handgun
	8669	Other skin graft to other sites
V-Codes	V5867	Long-term (current) use of insulin
	V1642	Family history of malignant neoplasm of prostate
	V1582	Personal history of tobacco use
	V462	Other dependence on machines, supplemental oxygen
	V8381	Cystic fibrosis gene carrier
	V8537	Body Mass Index 37.0-37.9, adult

Figure 7. Examples of ICD9-CM codes and long descriptions.

Our study used admission codes, which indicate the patient's original location before transfer to the hospital. In both the Maine Inpatient dataset and the Vermont Emergency Department dataset, these codes specify if an admitted patient came in through another hospital facility, the emergency room, law enforcement, etc. Similarly discharge codes indicate a patient's status including transferred to another facility, rehab, or deceased ("Expired"). We used these details to differentiate between candidate hospital records that match to the same named individual from the NewsStories.

Unlike diagnoses and procedure codes, admission and discharge status codes do not follow a standardized format across the states. Examples of coded numeric values for Maine and Vermont are listed in Figure 8.

Code Type	Maine		Vermont	
	Code	Description	Code	Description
Admission	1	Non-health care facility point of origin	1	Transfer from a hospital
	2	Clinic	2	Transfer from a skilled nursing facility
	7	Emergency room	7	Emergency room
	8	Court/law enforcement	8	Court/law enforcement
	C	Readmission to same home health agency	65	Newborn - born in this hospital
Discharge	20	Expired (Deceased)	1	To another acute hospital
	30	Still Patient	4	To another facility
	43	Discharged/Transferred To Federal Health Care Facility	5	Home - own or family care
	50	Hospice - Home	8	Died
	65	Discharged/Transferred To Psychiatric Hospital	13	Still a patient

Figure 8. Examples of admission and discharge status codes

Materials: NewsStories

We performed three searches on LexisNexis Academic, an aggregator and search engine for news stories, to find articles published in a sample of local newspapers in Maine and Vermont [25]. In Maine, our first search was for stories published in 2010 that contained the word "hospitalize," "hospitalized," or "transported," and retrieved 324 stories. Our second search used the terms "injured" and "hospital," and retrieved 316 stories. Finally, our search for stories containing "hospitalize" and "medical," "hospitalized" and "medical," "transported" and "medical," or "injured" and "medical" retrieved 472 stories. In total, our searches yielded 1,112 stories about car crashes, motorcycle accidents, ski accidents, and other incidents that resulted in hospitalizations in Maine hospitals. Using the same sequence of keyword searches for Vermont, we retrieved 201, 105, and 69 searches respectively, for a total of 375 news articles.

For Maine, the sample of newspaper stories came from local new sources including Bangor Daily News (Maine), Portland Press Herald (Maine), Brattleboro Reformer (Vermont), and Lowell Sun (Massachusetts). Vermont stories came from Brattleboro Reformer (Vermont), The Keene Sentinel (New Hampshire), and The Press-Republican (New York). The percentages in Figure 9 show that half of the stories in Maine came from the Bangor Daily News and more than half of the stories in Vermont came from the Brattleboro Reformer.

	Newspaper	News Stories	% of total stories
Maine	Bangor Daily News (ME)	538	48%
	Portland Press Herald (ME)	207	19%
	Foster's Daily Democrat (NH)	114	10%
	Portsmouth Herald (NH)	77	7%
	Brattleboro Reformer(VT)	22	2%
	Other	154	14%
	Total	1,112	100%
Vermont	Brattleboro Reformer (VT)	196	52%
	The Keene Sentinel (NH)	47	13%
	The Press-Republican (NY)	18	5%
	Other	114	30%
	Total	375	100%

Figure 9. List of newspaper sources that publish stories about local accidents and hospitalizations.

News stories about the same incident often appeared more than once in the LexisNexis results. Some stories described patients sent to hospitals in neighboring states. A few stories concerned people who died before ever being admitted to a hospital. Other stories lacked the name of the individual. After we removed duplicates and stories lacking the requisite name of a patient to a Maine or Vermont hospital, we had a sample of 177 stories in Maine and 38 stories in Vermont. We term these as the "MaineNewsStories" and the "VermontNewsStories." These collections of text include the newspaper, byline, publication date, and content of each news story.

Subjects

In Maine and Vermont, the news stories referenced individuals involved in a motor vehicle accident, a fire, or other accidents resulting in a visit to the hospital. These incidents from news stories that appeared in Maine or Vermont newspapers also contained demographic details about the individual, such as age and hometown, the incident that led to the hospitalization, and the hospital involved.

Approach

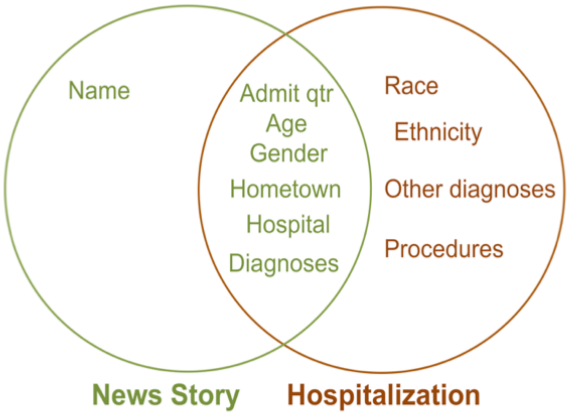


Figure 10. Re-identification strategy to associate a named individual in a news story to a de-identified hospital record. The match uses unique occurrences of the combination of age, gender, hometown, hospital, admission quarter, and diagnoses, as related to the incident described in the news story.

The approach is simple: match information from each news story describing a named individual's visit to a hospital, to the statewide data of hospital visits and see which matches uniquely associate one person from a news story to one and only one hospital record.

Figure 10 shows a graphical depiction. A news story contains the patient's name, age, gender and hometown, identifies the hospital, the patient's admission date, and describes the incident in ways that relate to diagnosis codes. If patient information from a news story matches one and only one hospitalization record, then we say it is a "unique match" (or re-identification). A match associates a named individual to a distinct hospital record.

This is the same as approach used in the Washington State experiment [1]. We match patient demographics and incident descriptions found in news stories to hospital records. We followed the steps below to conduct the matches for both Maine and Vermont.

Step 1. Transcribe News Stories

In Step 1, we read each news story in the compilation of 1,112 MaineNewsStories and 374 VermontNewsStories. We recorded the name, age, hometown of the patient, publication date, hospitalization and injury details extracted from the newspaper stories, and compiled these as the "NewsData." We did not transcribe the hospitalization details or demographic record of any individual who was not mentioned by name. If an individual's name appeared in more than one news story, we deleted any duplicates and only counted an individual's name twice if he or she was involved in another incident that resulted in a hospitalization. Some NewStories do not name injured individuals because reporters may not be able to verify the name or choose to redact name of a minor for legal reasons.

Step 2. Direct Matching of NewsData to Hospitalizations

We matched information from each news story describing a named individual's hospital visit to the statewide data. In the end, we determined which matches uniquely associate one person from a news story to one and only one hospital record.

Step 2a. In the first phase of the matchup, we wrote a simple computer program that matched the information extracted from the news stories (NewsData) to the de-identified hospital records (MaineDataset) using the following five fields: Hospital Code (hp), Geocode (geocode), Gender (sx), Age (age), and Admission Quarter (admqtr). To match the hospital records in the VermontDataset, we used Hospital Number (HNUM2), ZIP code group (ZIPGRP), Gender (SEX), Age Group (AgeGRP) and Admission Quarter (ADMID_QTR).

Step 2b. In the second phase, we took the subset of named individuals who matched to at least one de-identified hospital record based on hospital code, geocode, gender, age and admit quarter. We conducted a thorough visual verification that confirmed that the injury from the NewsData matched with additional details from the hospital. We inspect all available data points including: diagnosis codes (dx1–dx9), procedure codes(ps1–ps10), length of stay (los), point of origin of admission (sadm), discharge status (disp), race, and ethnicity. The VermontDataset does not contain procedure codes or race and ethnicity.

Finally, for the MaineDataset we inspected the procedure dates (MM/DD/YY) and validate whether they matched the time period of the accident according to the newspaper article. If there was no procedure date listed, we did not use procedure date as a criterion for matching. The VermontDataset does not have procedure codes or dates.

Step 3. Identifiability of the Data

Sweeney previously found that the "identifiability," or the number of records that share the same set of characteristics within a dataset, of the HIPAA Safe Harbor provision was 0.04 percent [26][27]. That is, 0.04 percent of records are uniquely identifiable based on only {year of birth, gender and ZIP code}.

We assessed the baseline level of how many individuals were already uniquely identifiable in each dataset to better understand re-identification results. For the MaineDataset, we calculated identifiability using {age, gender, geocode, hospital and admission quarter}. In the VermontDataset, we used {age group, gender, ZIP code group, hospital and admission quarter}. Then, we evaluated the level of identifiability again without the admission quarter.

Step 4. Direct Matching using a HIPAA Compliant Version of the Data

Neither the MaineDataset and VermontDataset meet the minimum HIPAA Safe Harbor requirements, as both provide more details with dates and geographical identifiers than Safe Harbor allows. Maine and Vermont share quarter, which is more specific than providing dates in years. They also share geographic identifiers, such as geocode or ZIP code. We constructed "HIPAA Datasets" for each statewide dataset to test the effectiveness of HIPAA Safe Harbor guidelines.

For the MaineDataset, we redacted all geocodes that had a 2010 population of less than 20,000. See Figure 11 for remaining geocodes. We redacted all dates not listed as years, including admission quarter and procedure dates. We call this new, redacted data the "Maine HIPAA Dataset."

Geocode	Town Name	Population in 2010	Redacted 3-digit geocode
5170	Portland	66,194	051
1050	Lewiston	36,592	010
19020	Bangor	33,039	190
1010	Auburn	23,055	010
31050	Biddeford	21,277	310
31230	Sanford	20,798	312

Geocode	Town Name	Population in 2010	Redacted 3-digit geocode
5220	South Portland	25,002	052

Figure 11. List of geocodes that remains in Maine HIPAA Dataset.

There is an aggregate "ZIP code group" field in the VermontDataset for patient location information. However, some ZIP code groups do not have a population over 20,000. We redact all groups of ZIP codes that had a 2010 population of less than 20,000. See Figure 12 for remaining ZIP code groups. We redacted the admission quarter since it is a date indicator that is not listed in years. We call this redacted data the "Vermont HIPAA Dataset."

ZIP Code Group	Description	Population in 2010	Redacted 3-digit ZIP
37	ZIP in 03700-03799 range	77,452	037
50	ZIP in 05000-05099 range	61,355	050
51	ZIP in 05100-05199 range	29,595	051
53	ZIP in 05300-05399 range, excluding 05301	20,458	053
54	ZIP in 05400-05499 range, excluding Burlington to Saint Albans	125,682	054
56	ZIP in 05600-05699 range, excluding Barre and Montpelier	56,237	056
57	ZIP in 05700-05799 range, excluding 05701	61,686	057
128	ZIP in 12800-12899 range	225,336	128
129	ZIP in 12900-12999 range	164,172	129
136	ZIP in 13600-13699 range	236,059	136
999	Non-USA or Unknown ZIP		999
5401	Burlington	28,185	054
5701	Rutland	20,802	057
034+036	ZIP in 03400-03499 or 03600-03699 range	97,678	034+036
058-059	ZIP between 05800 and 05999	62,345	058-059
120-123	ZIP in 12000-12399 range	924,965	120-123

ZIP Code Group	Description	Population in 2010	Redacted 3-digit ZIP
OTH-MA	All MA ZIP codes		OTH-MA
OTH-NH	All other NH ZIP codes		OTH-NH
OTH-NY	All other NY ZIP codes		OTH-NY
OTH-ST	All ZIP codes other than VT, NH, NY, and MA		OTH-ST

Figure 12. List of ZIP code groups that remains in Vermont HIPAA Dataset.

We wrote a computer program to match named individuals to de-identified hospital records using age, gender, hometown and hospital code. Then, we systematically reviewed any named individual who matched at least one de-identified hospital record. The goal was to verify the injury with a de-identified record using additional data points. In Maine, this includes: diagnosis codes, procedure codes and dates, admission and discharge codes, race and ethnicity, and length of stay, if available. In Vermont, we excluded procedure codes and dates, and race and ethnicity.

We note that the unique matches in Washington State's data were all correct, as verified by contacting the re-identified patients. Since the methods in this paper reproduce the basic methodology of matching news stories with de-identified hospital records, we assumed a similar level of correctness and did not contact patients to verify the matches. While the quality of the news stories in Maine and Vermont may vary from those in Washington, this study assumes that the level of detail in local reporting does not vary since the detail in news stories from Maine and Vermont are similar to the hospitalization details in the Washington news articles.

Results

For clarity, we report our findings following the same steps that are used to describe our methods.

Results for Step 1. Transcribe News Stories

There were 177 news articles that mentioned 244 named individuals taken to a Maine hospital. The subjects in Vermont are a sample of 47 named individuals mentioned in the 38 news stories who required hospitalization in Vermont. We only counted news articles that name at least one individual who was admitted to a Maine hospital in 2010 or who visited a

Vermont emergency room in 2011. Figures 13 and 14 show the fields present in news stories for named individuals in each state.

Number of Fields	Name	Gender	Age	Geocode	Hospital	At Least 1 Injury Detail	Number of Subjects	Totals
4	x	x			x		9	9
5	x	x	x		x	x	11	28
5	x	x	x	x		x	11	
5	x	x		x	x	x	6	
6	x	x	x	x	x	x	207	207
						Total	244	244

Figure 13. Distribution of fields in Maine NewsData.

Number of Fields	Name	Gender	Age	Geocode	Hospital	At Least 1 Injury Detail	Number of Subjects	Totals
4	x	x			x	x	2	4
4	x	x		x	x		2	
5	x	x		x	x	x	3	6
5	x	x	x		x	x	2	
5	x	x	x	x	x		1	
6	x	x	x	x	x	x	36	36
						Total	47	47

Figure 14. Distribution of fields in Vermont NewsData

Results for Step 2. Direct Matching of Newspaper Stories to Hospitalizations

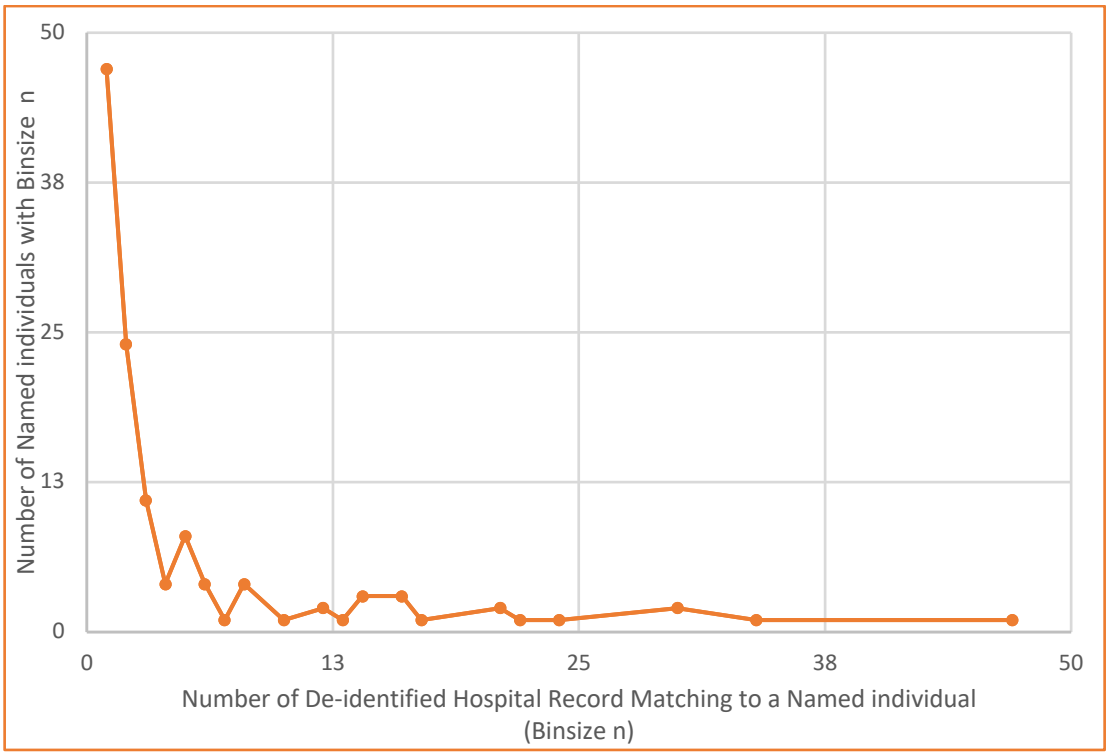
Results for Step 2a. We matched each of the named records in NewsData with the hospital records using only age, gender, geocode, admission quarter, and hospital code. In Maine, 136 of the 244 named individuals (or 55.5 percent) matched with at least one hospital record (binsize k>1) and 47 out of the 244 (or 19.3 percent) matched with exactly one de-identified hospital record (binsize k=1).

Out of the 136 named individuals, 24 (or 17.6 percent) had two candidate hospital records (binsize 2) and 11 (or 8.1 percent) that had three candidate hospital records (binsize 3). In sum, 89 (or 65.4 percent) named individuals had more than one candidate hospital record. Figure 15 shows the breakdown of how many named individuals matched to *n* de-identified records by binsize. A total of 114 out of 136 (or 83.8 percent) named individuals had re-identifications of binsize k<20 and 94 of the 136 (or 69.1 percent) named individuals had a binsize k≤5.

Number of De-identified Records Matching to a Named individual (Binsize)	Number of Named individuals with Binsize n
1	47
2	24
3	11
4	4
5	8
6	4
7	1
8	4
10	1
12	2
13	1
14	3
16	3
17	1
21	2
22	1
24	1
30	2

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901

Number of De-identified Records Matching to a Named individual (Binsize)	Number of Named individuals with Binsize n
34	1
47	1



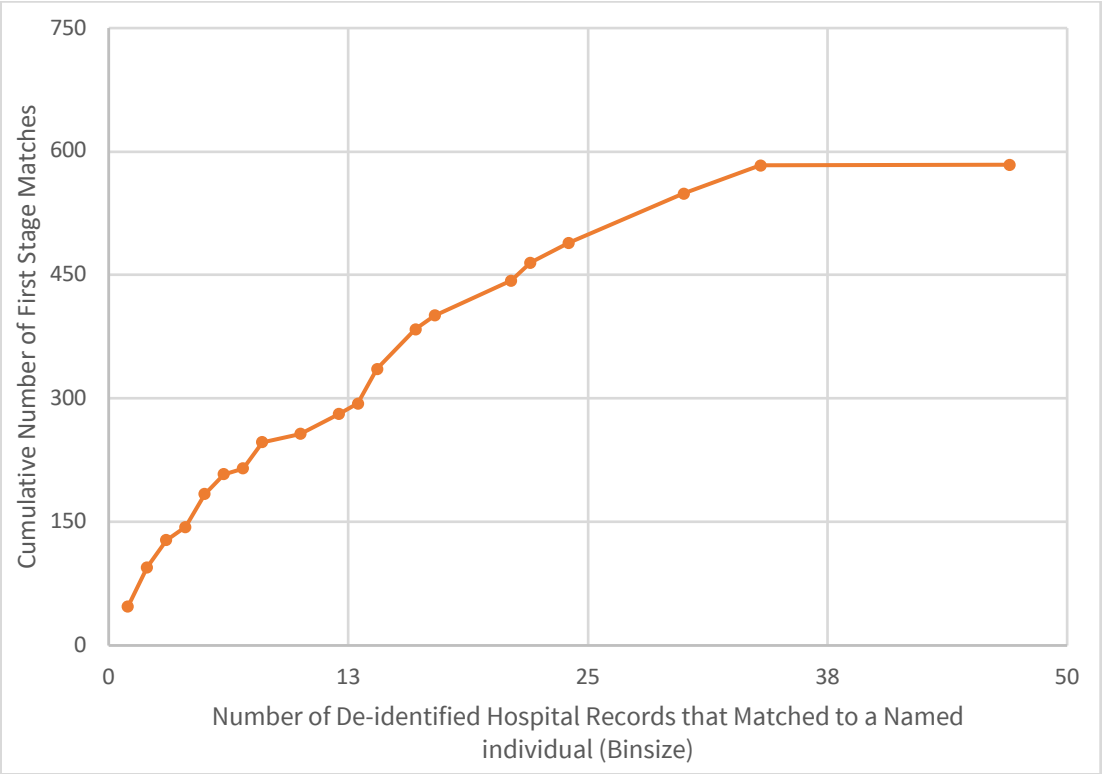


Figure 15. Matching records where k<50 are graphically sorted by binsize. First stage re-identifications are based only on age, gender, geocode, admit quarter and hospital code. All demographics came from Maine newspaper articles. In Step 2b, we matched these named individuals to a de-identified hospital record based on diagnoses code, discharge code, procedure code and date, length of stay, race, and ethnicity, if available. The cumulative number of named individuals matching to a de-identified hospital record are represented as well.

In the first stage of the matchup for Vermont, we ran the python code to cross-reference Vermont NewsData with the VermontDataset based on age group, gender, ZIP code group, admission quarter, and hospital code. Forty out of the 47 (or 85.1 percent) named individuals matched to at least one de-identified hospital record (binsize k≥1). Two out of the 40 matched to exactly one hospital record and 39 named individuals matched to more than one record. We graphically showed how many de-identified records matched to each named individual using binsizes where k<50 in Figure 16. We cross-linked these named individuals in Step 2b using diagnoses code, discharge code, procedure code and date, and length of stay.

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901

Number of De-identified Records Matching to a Named individual (Binsize)	Number of Named individuals with Binsize n
1	2
2	3
3	1
5	3
9	1
10	2
12	2
14	1
15	1
16	1
18	1
22	1
24	1
45	2

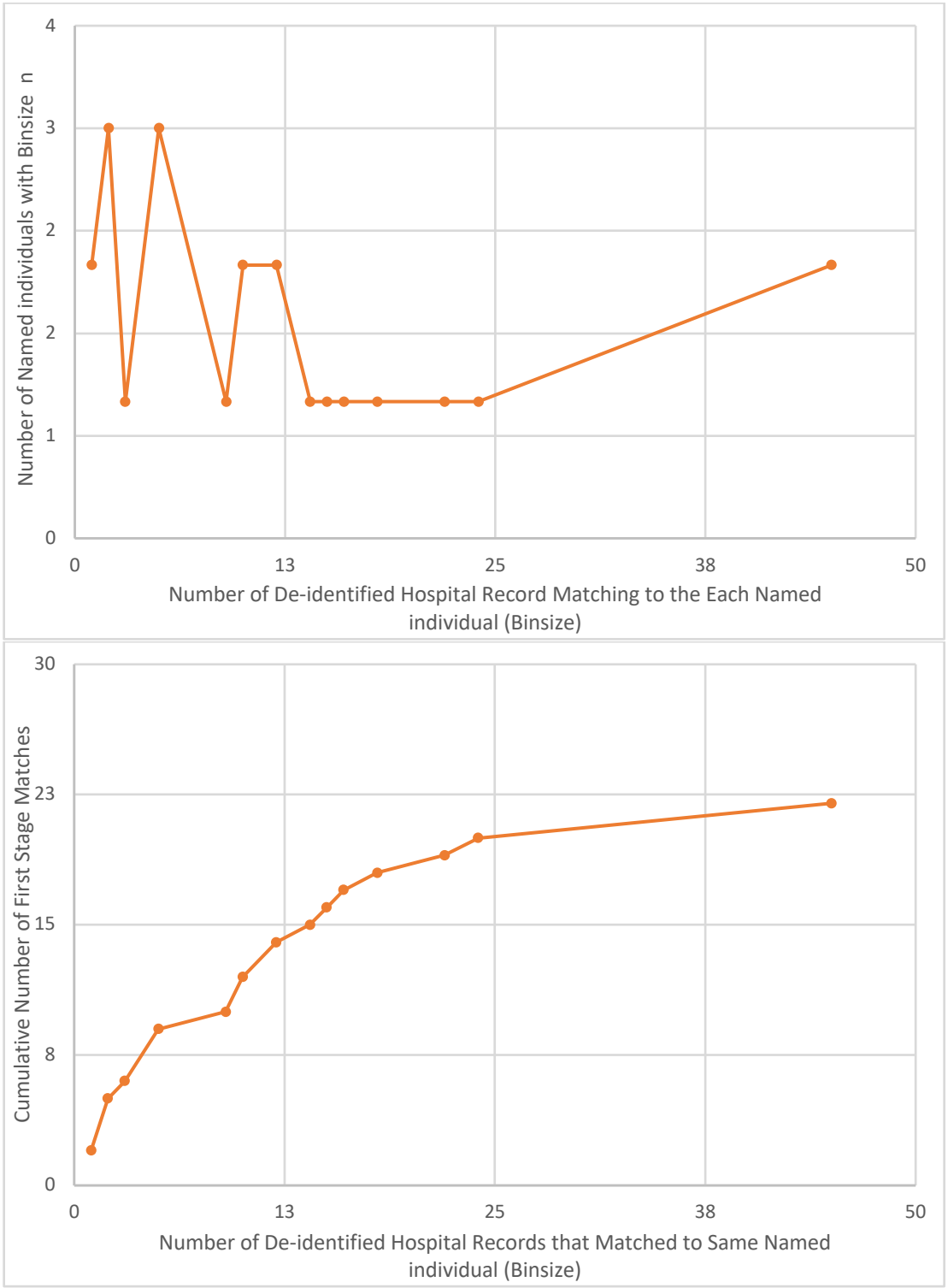


Figure 16. VermontDataset re-identifications based only on age group, gender, ZIP code group, admission quarter and hospital code. Matching records with binsize k<50 are graphically displayed by binsize.

Results for Step 2b. After Step 2a, there were 136 named individuals who matched to one or more hospital records in the MaineDataset after matching the de-identified records based on age, gender, geocode, admission quarter, and hospital code. We confirmed that the diagnoses codes, discharge code, procedure date, length of stay, race, and ethnicity from the de-identified record matched the specifications from the news data if the new story included the data.

Out of the 47 named individuals that matched exactly to one de-identified record in Maine, 26 uniquely matched a de-identified hostpital record after visual verification. Out of the 89 named individuals that matched to more than one de-identified hospital record, 35 uniquely matched to one de-identified record after visual verification. In total, we found 69 out of 244 (or 28.3 percent) unique re-identifications. Figure 17 describes the process of elimination based on this methodology.

In 2010, 94.4 percent of the population in Maine was White. The MaineDataset was also homogenous: 95.4 percent of the records referred to White patients. When verifying race and ethnicity, we could more easily check whether minorities in the de-identified hospital data matched our newspaper stories because we could use online searches for public images (if available) to confirm a match.

There were 47 named individuals that matched to one or more hospital records in the VermontDataset based only on age group, gender, ZIP code group, hospital, and admission quarter. After verifying diagnoses code, discharge status, and length of stay, 16 named individuals matched to a single hospital record. Figure 17 visually depicts the re-identification strategy for Vermont.

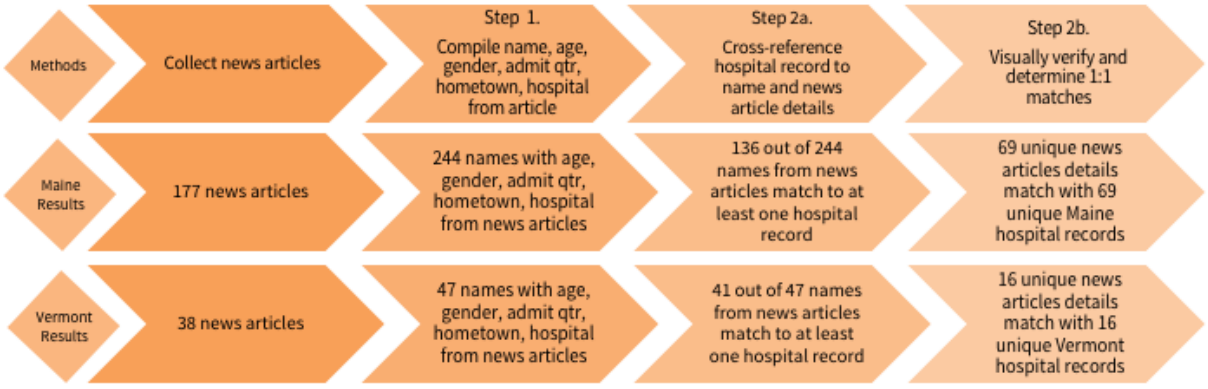


Figure 17. Outline of each stage of the re-identification strategy. There were 69 out of 244 (or 28.3 percent) unique re-identifications in Maine and 16 out of 47 (or 10.5 percent) unique re-identifications in Vermont. In Step 2a, we cross-referenced NewsData (hometown, age, gender, hospital, and admission quarter) with hospital data (geocode or ZIP code group, age or age group, gender, hospital, admission quarter). In

Step 2b, we verify the records using diagnoses, discharge, admission codes, length of stay and if available, procedure date, race, and ethnicity.

Results for Step 3. General vulnerability based on the identifiability of the data

Previous studies found identifiability of data cleaned under HIPAA Safe Harbor provision to be 0.04 percent [26][27]. In comparison, there are 56,735 records out of 105,808 (or 53.6 percent) that are uniquely identifiable using the characteristics {age, gender, geocode, hospital and quarter} in the MaineDataset and 4,570 records out of 268,984 (or 0.55 percent) in the VermontDataset using the same combination of fields. We denote unique identifiability when there is only one record in a de-identified dataset with the specific combination of field values mentioned above. A record with a binsize of 2 indicates that there are 2 records in the de-identified dataset that share the same specific combination of values. The higher the number of records that are uniquely identifiable in the dataset, the higher the identifiability of the dataset.

More than half of the records in the MaineDataset are unique when only considering {age, gender, geocode, hospital and quarter}. Figure 18 shows the binsize results. The minimum binsize is 1, the maximum binsize is 70, the average binsize is 24 and standard deviation of 17. Records having a binsize of 1 and 2 make up 73.7 percent (or 78,027 out of 105,808) of the MaineDataset, and records that have binsizes 1 through 3 make up 83.4 percent of all de-identified records.

When we consider the identifiability of the MaineDataset excluding admission quarter, the percentage of values that are unique decreases to 33.7 percent and binsize total increases to 63. The minimum binsize is still 1, but maximum binsize increases to 231, with an average binsize of 43 and standard deviation of 41. The cumulative records that have binsize of 1 and 2 make up 51.3 percent of the MaineDataset. Binsizes of 1 through 3 make up 61.4 percent of the records. The exclusion of admission quarter decreases the number of unique values (binsize 1) and increases the binsize of the de-identified records. Together, Figures 18 and 19 show a comparison of the larger binsizes and lower number of unique records when quarter is excluded.

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901

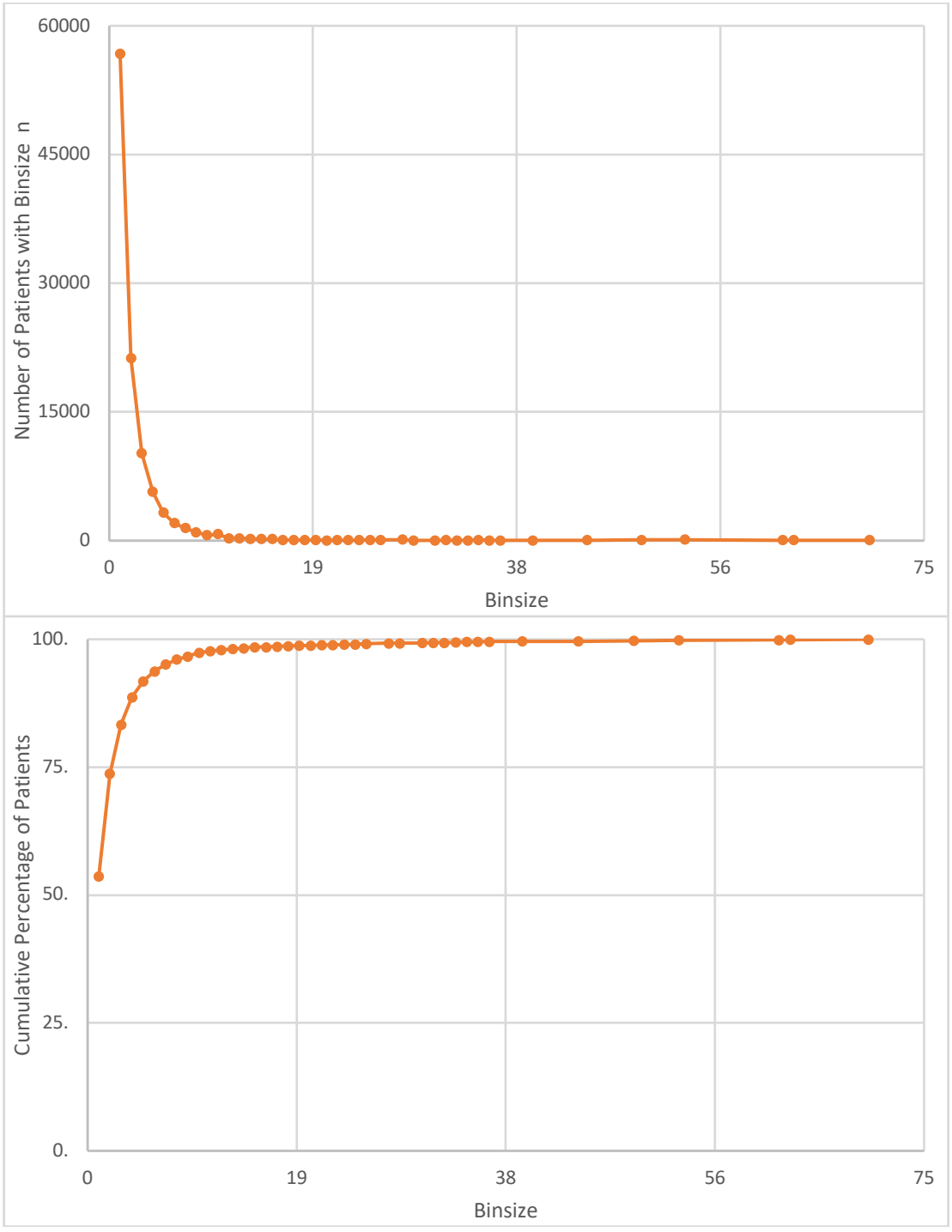
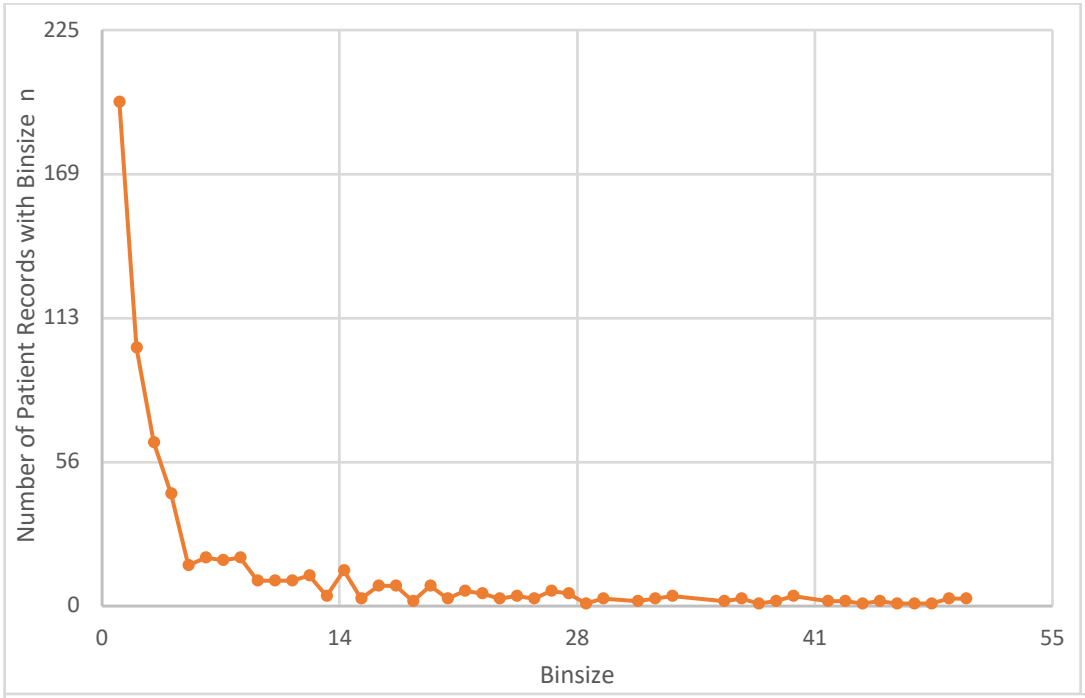
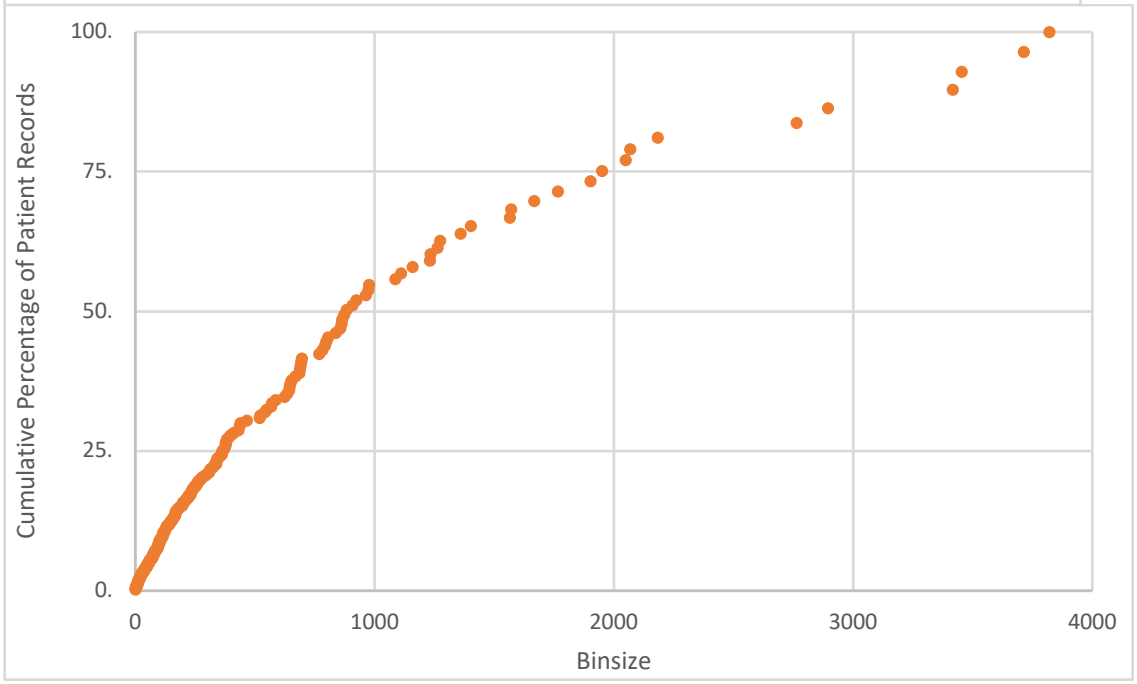


Figure 18. MaineDataset identifiability of records using of {age, gender, geocode, hospital, quarter}. Unique records that have binsize 1 make up more than half of the total number of de-identified records. The first chart shows the counts by binsize for all binsizes (range 1 to 70) and the second chart shows identifiability rates by cumulative percentage for all binsizes.

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901





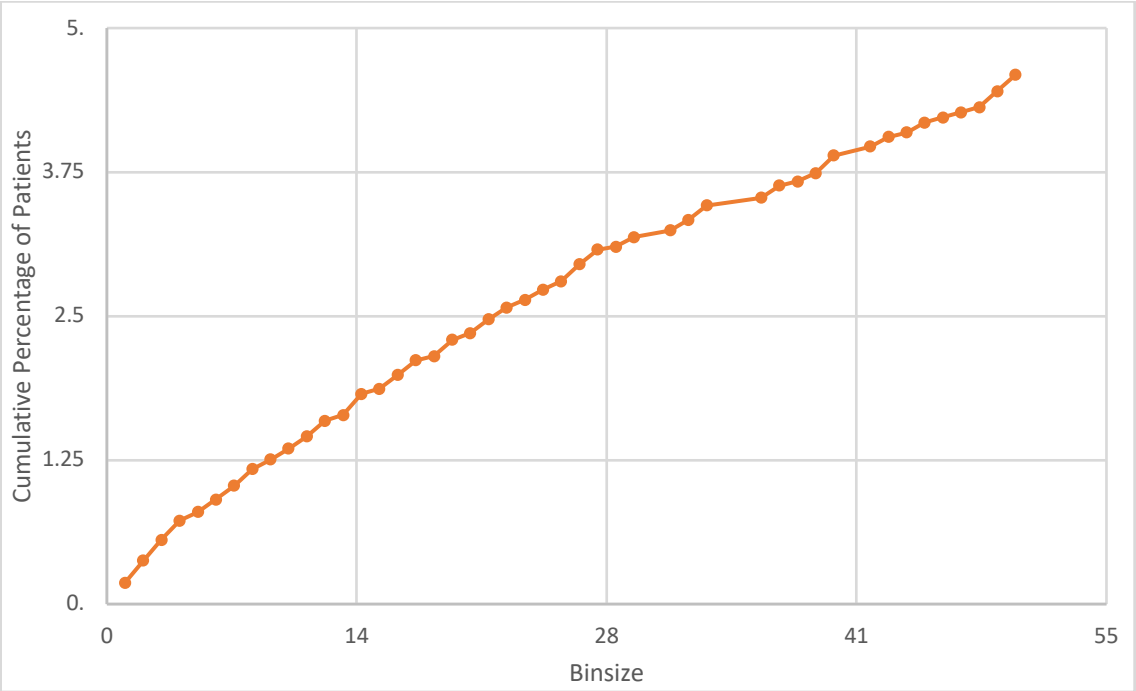
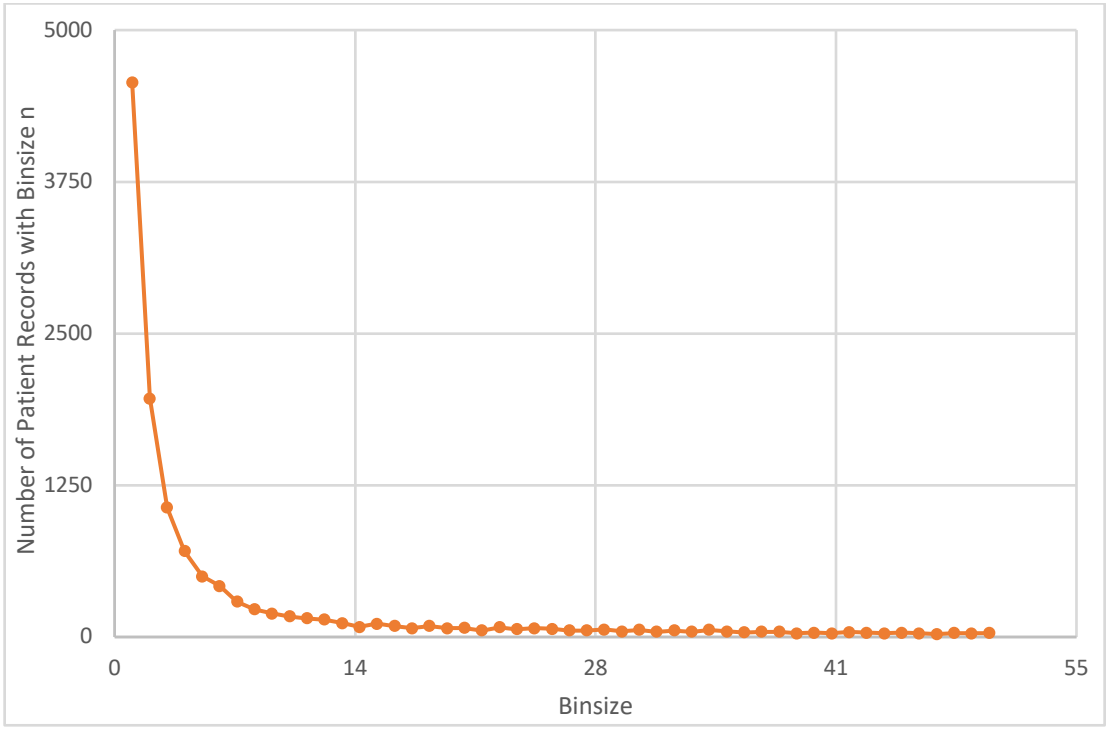
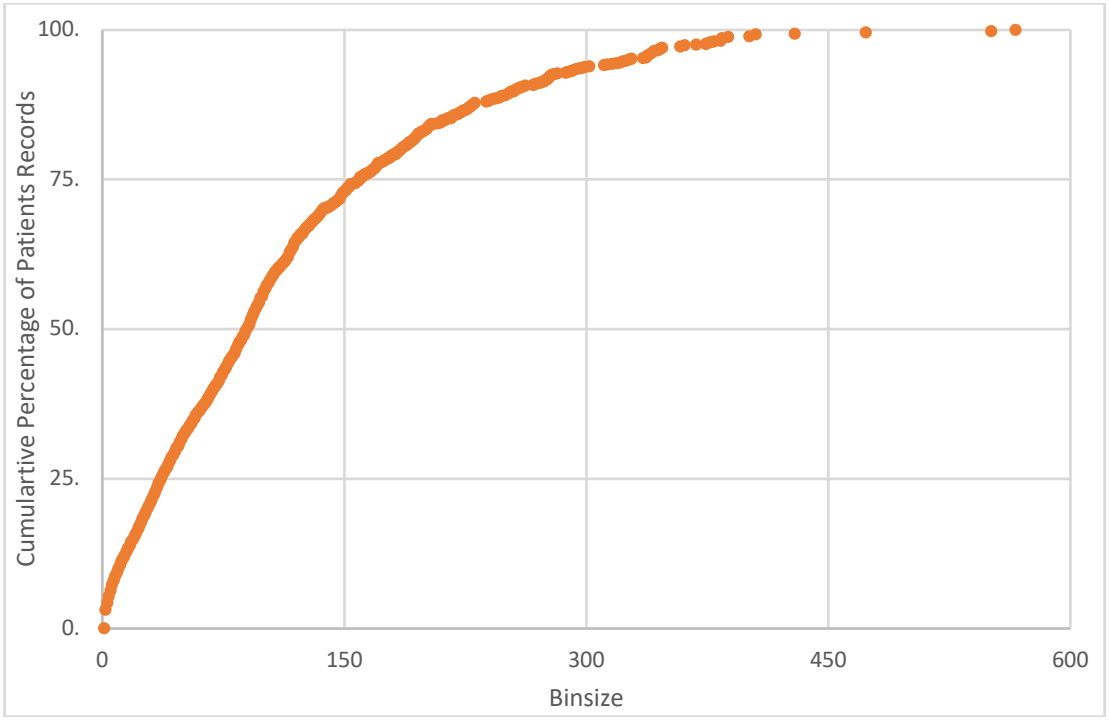


Figure 19. MaineDataset identifiability of records using of {age, gender, geocode, hospital}. The first chart shows the number of patient records that have binsizes that are 50 or less. The second chart displays identifiability rates by binsize. The third chart zooms in on the second chart and shows the cumulative identifiability rates by binsize for the binsizes 1 to 50.

The VermontDataset has fewer unique records. This is due in part to Vermont's reporting of age and ZIP code using age group and ZIP code group. Out of 268,984 records, there were 4,570 records (or 1.7 percent) identifiable to a binsize of 1 and 8,501 (or 3.2 percent) records identifiable to a binsize of 1 and 2. There are a total of 301 binsizes. These are significantly lower identifiability estimates than the MaineDataset. This result is expected because the VermontDataSet aggregates both age and ZIP code into groups. Subsequent charts show the cumulative percentage of patients according to binsize and the high percentage of patients that have a small binsizes. Minimum binsize is 1, maximum binsize was 566 with average binsize of 164 and standard deviation of 109. Figure 20 depicts the identifiability of records in the VermontDataset.

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901





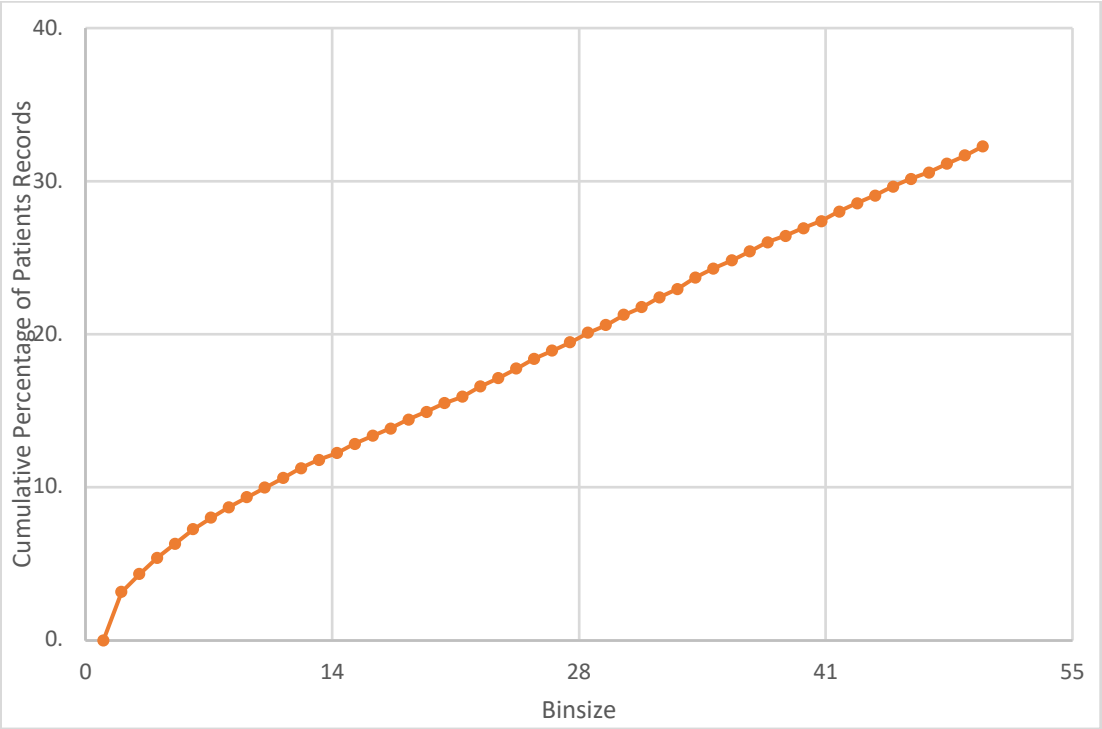
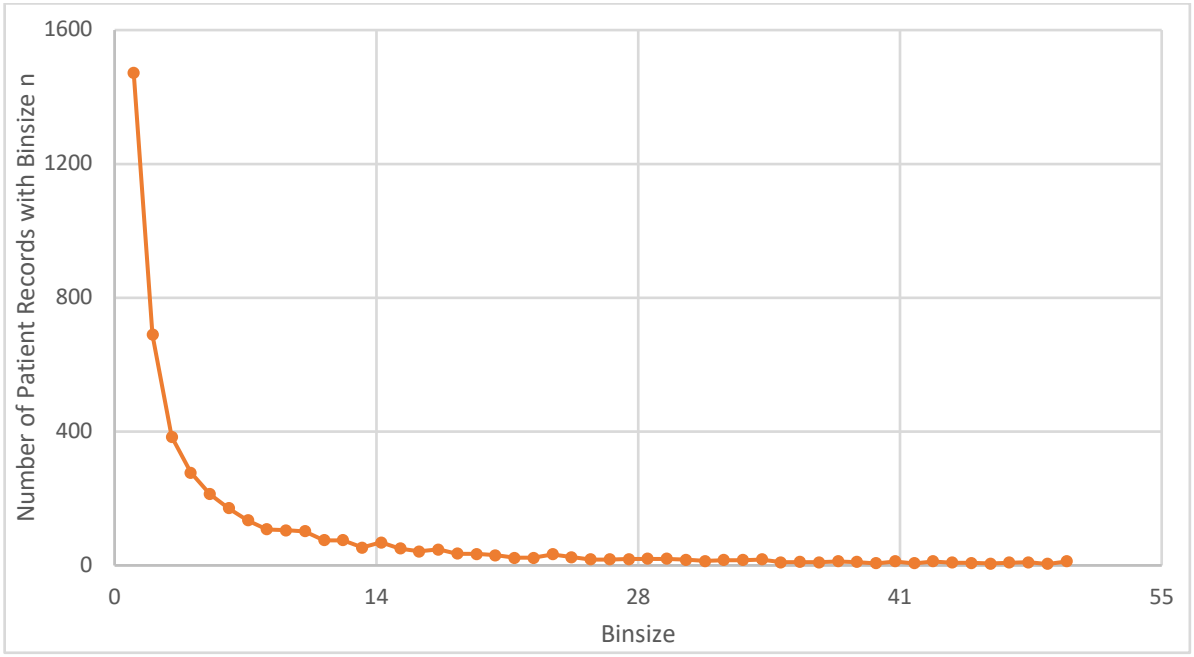
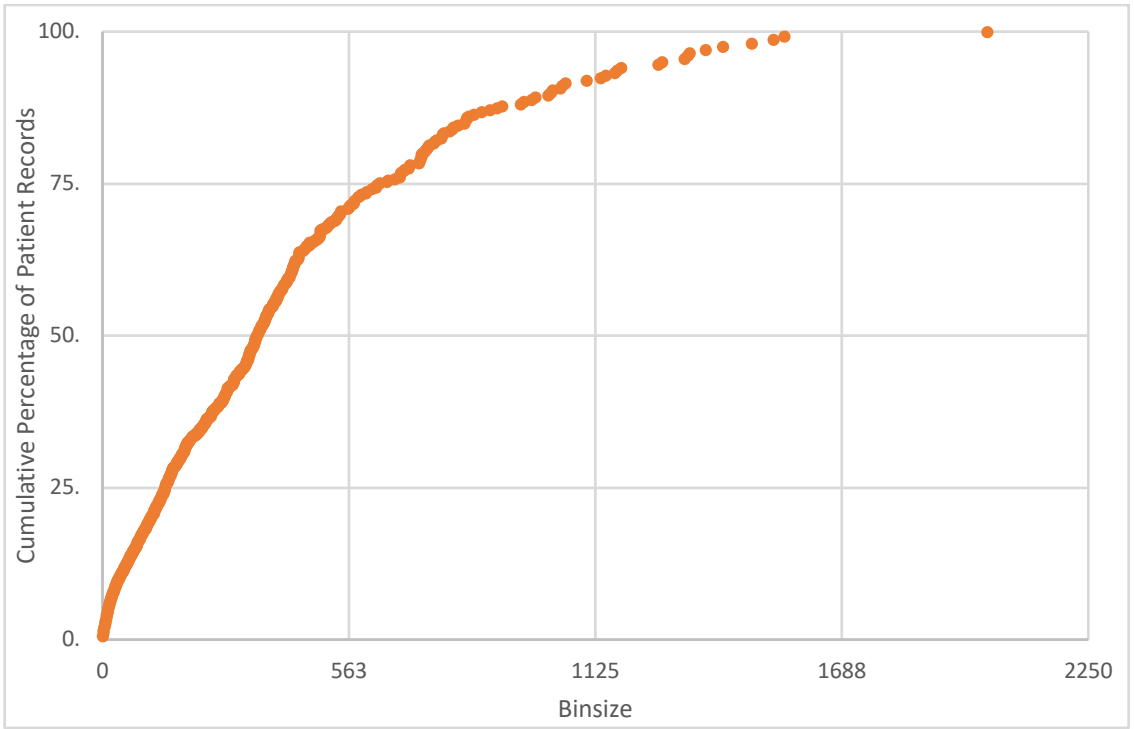


Figure 20. Identifiability of records in the VermontDataset based on {age, gender, geocode, hospital and quarter}. The first graph shows the number of records with binsize of 50 or less (k≤50) in order to show the number of unique records already in the data. The second graph shows the cumulative percentage of patients who are identifiable to show the holistic binsize range and identifiability relative to the entire VermontDataset. The third graph zooms in on the second graph and shows the identifiability rates for the first 50 binsizes.

After we exclude admission quarter, the minimum binsize is 1, maximum is 2,019, with an average of 340 and standard deviation of 308. There are 1,473 records (or 0.55 percent), 2,851 (or 1.06 percent) with a binsize of 1 and 2, and 4,003 (or 1.49 percent) with binsize of 1 through 3. The total number of binsizes is 465. Again, unique records decrease when we exclude admission quarter. Figure 21 shows the frequency of binsizes that are less than 50.

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901





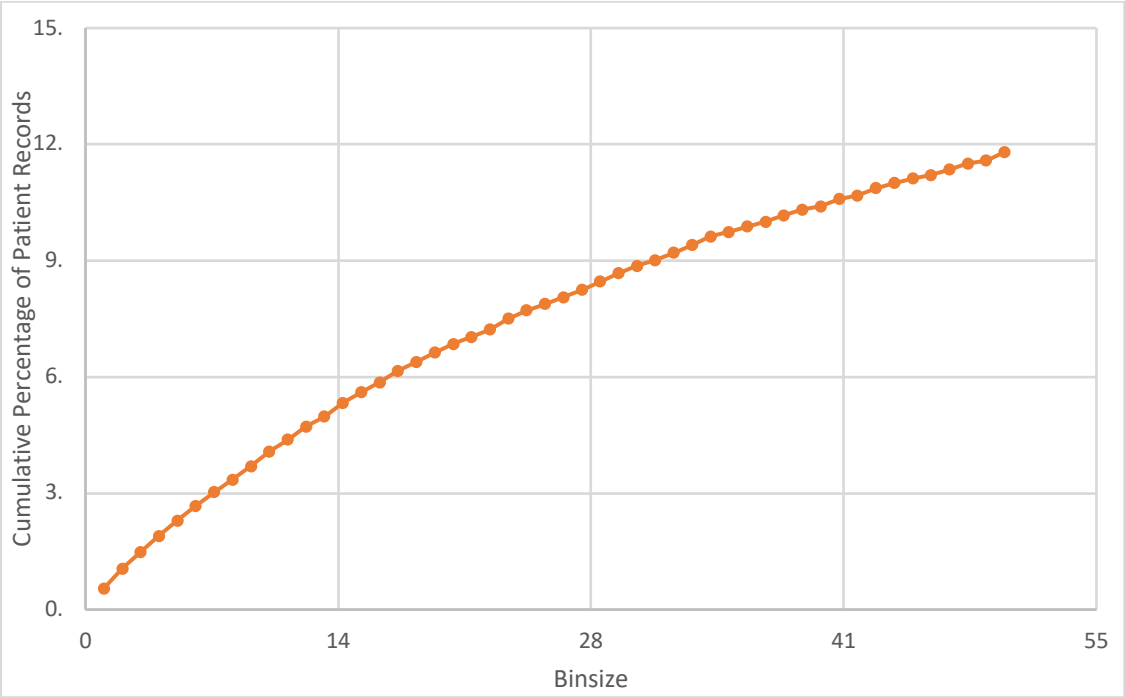


Figure 21. Identifiability of records based on {age group, gender, ZIP code group, and hospital} in the VermontDataset. The first graph shows the number of records with binsize of 50 or less (k≤50) and the second graph shows the cumulative percentage of identifiable patients by binsize. The third graph zooms in on the second graph and shows the identifiability rates for the first 50 binsizes.

Even though the number of records that have binsize of 1 within the VermontDataset is lower than those in the MaineDataset, in Step 3 we observed a higher rate of re-identification for Vermont than for Maine. This might be explained by the fact that Vermont and Maine contain different kinds of hospitalization data. The VermontDataset is a record of all emergency department visits and MaineDataset is a record of inpatient admissions. Not all the named individuals in the news articles for Maine were reported as being "admitted" to a hospital after a motor vehicle accident, fire, or incident, so it is more likely that named individuals could be found in an emergency department dataset than in an inpatient dataset when they are "transported" or "seen" in the hospital.

Results for Step 5. Direct Matching using a HIPAA Compliant Version of the Data

We created a hospital dataset that followed the minimum HIPAA Safe Harbor guidelines. We redacted admission quarter and all geocodes from the MaineDataset that had a 2010 population of fewer than 20,000. There were 694 geocodes present in the MaineDataset, but only 7 remain present in the Maine HIPAA Dataset: Portland, Lewiston, Bangor, Auburn, Biddeford, Sanford, and South Portland (See Figure 11).

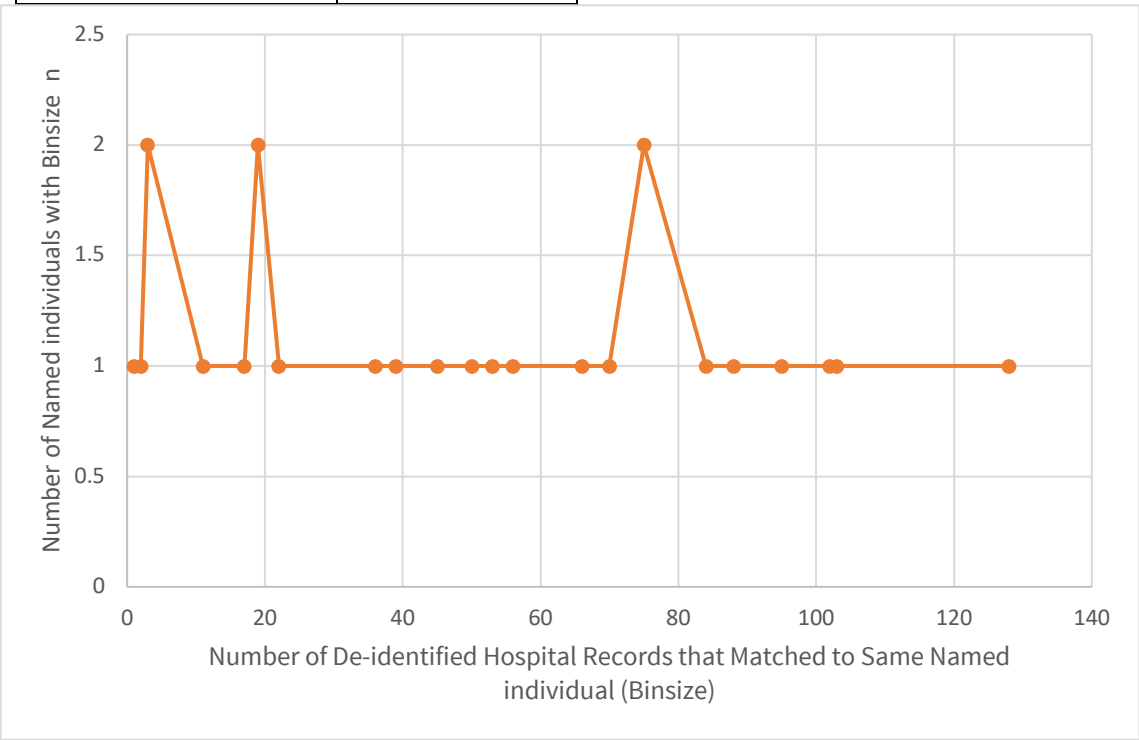
We did the matchup only on Hospital Code (hp), Gender (sx), and Age (age). There were 40 named individuals who matched to multiple de-identified hospital records of a binsize greater than or equal to 1. The distribution of the matchup is listed in Figure 22 according to binsize. There were fewer named individuals with small binsizes in the HIPAA Maine results since there is less specificity when 99.14 percent (688 out of 694) of geocodes were omitted from the Maine HIPAA Dataset. The minimum binsize was 1, maximum was 10,433, and average binsize was 782 with a standard deviation of 2,213.

We took the 40 potential matches and verified diagnoses codes, discharge codes, procedure codes, and length of stay, if available. We uniquely re-identified eight out of 244 (or 3.28 percent) HIPAA-compliant records from the MaineDataset.

Number of HIPAA De-identified Records Matching to the Same Named individual (Binsize)	Number of Named individuals with Binsize n
1	1
2	1
3	2
11	1
17	1
19	2
22	1
36	1
39	1
45	1
50	1
53	1
56	1
66	1
70	1

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901

Number of HIPAA De-identified Records Matching to the Same Named individual (Binsize)	Number of Named individuals with Binsize n
75	2
84	1
88	1
95	1
102	1
103	1
128	1
190	1



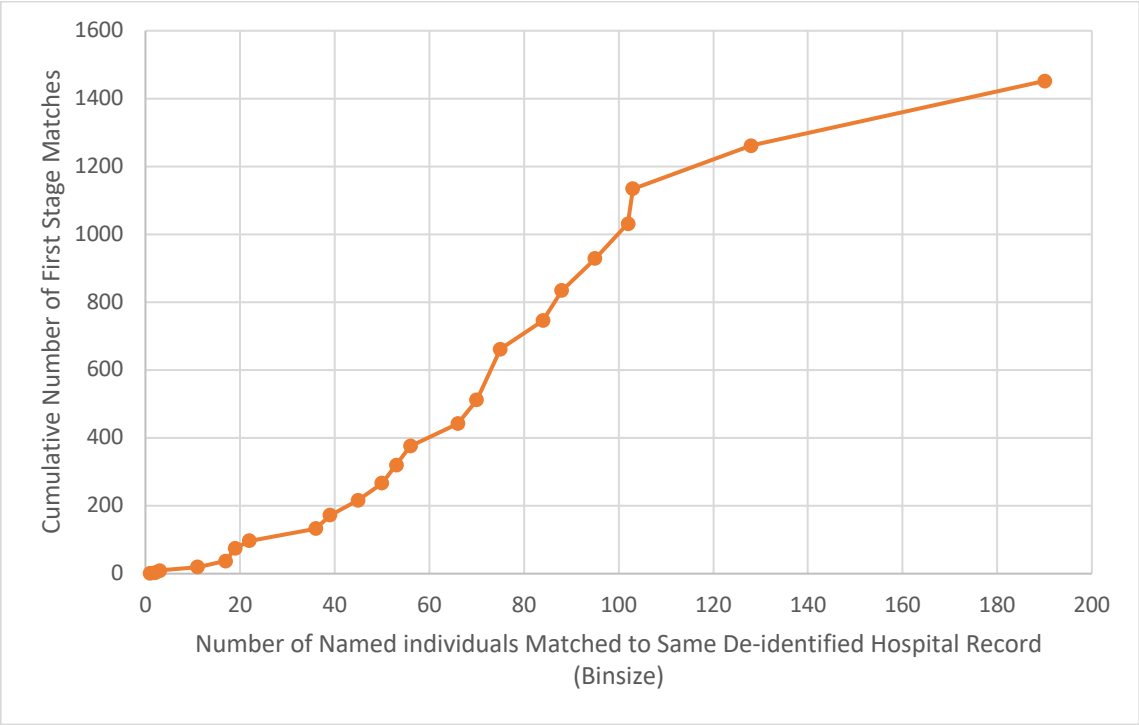


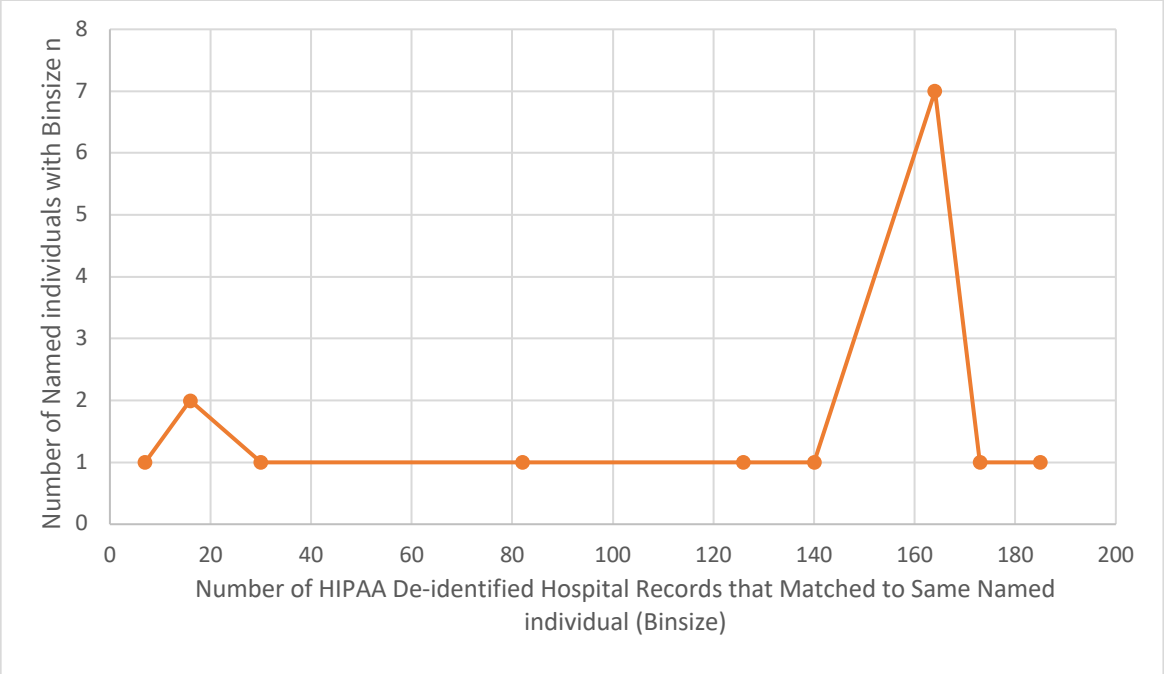
Figure 22. First stage re-identifications of HIPAA-compliant de-identified hospital records based only on age, gender, geocode (if population>20,000), and hospital code. We matched these candidates' de-identified hospital records further. We did not consider records that matched to more than 200 named individuals (k>200) for further visual inspection. We sorted records with binsize > 200 graphically and show the cumulative number of de-identified hospital records by binsize.

The VermontDataset does not follow HIPAA Safe Harbor standards. We removed any ZIP code groups with populations of fewer than 20,000 to create the Vermont HIPAA Dataset. In total, we dropped 10 ZIP code groups from the VermontDataset to create the HIPAA-compliant dataset. The ZIP code groups left in the de-identified dataset are listed in Figure 12. Forty-four named individuals match to at least one de-identified HIPAA hospital record. Again, binsize indicates the number of de-identified HIPAA records that match to a named individual. The minimum binsize was 7, maximum was 9,297, average binsize was 1,503 with a standard deviation of 2,595. See Figure 23 for binsize frequencies.

We took the matches that had a binsize less than or equal to 200 and further verified potential matches using diagnosis codes, discharge codes, procedure codes, and length of stay, if available. We uniquely re-identified five out of 47 (or 10.6 percent) HIPAA-compliant records from the VermontDataset.

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901

Number of HIPAA De-identified Records Matching to the Same Named individual (Binsize)	Number of Named individuals with Binsize n
7	1
16	2
30	1
82	1
126	1
140	1
164	7
173	1
185	1



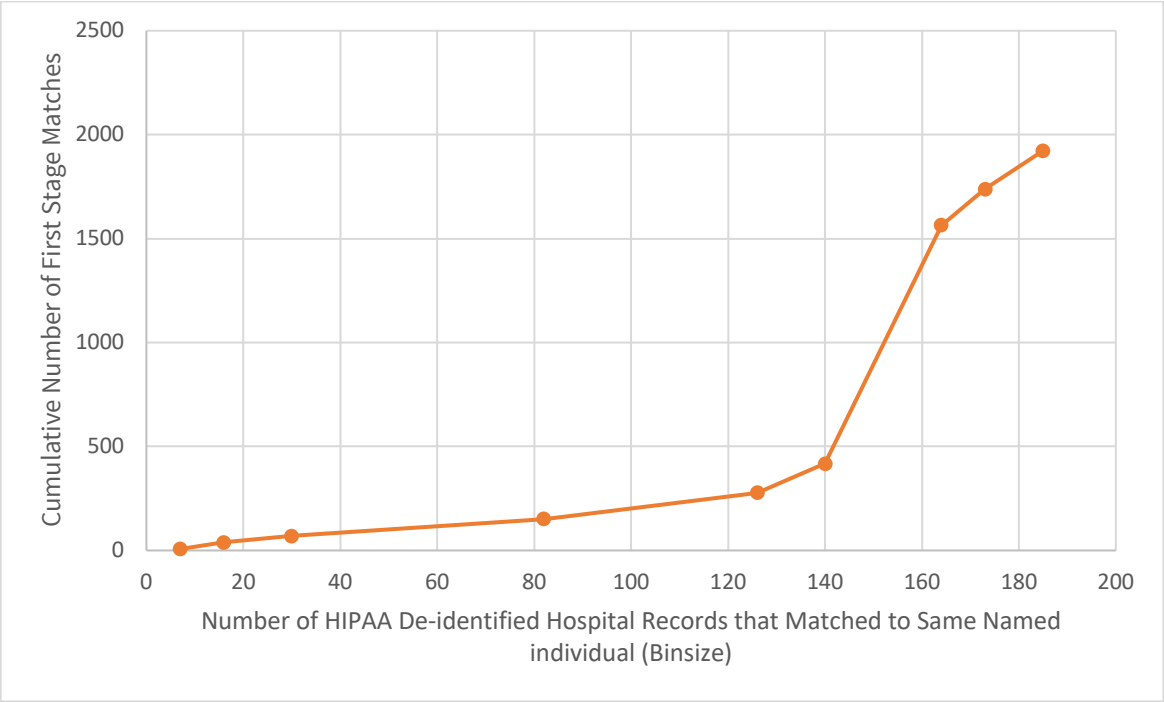


Figure 23. Vermont HIPAA Dataset re-identifications by binsize. These named individuals later match to a single HIPAA de-identified hospital record using visual inspection. Records that matched to more than 200 named individuals (k>200) were not considered for further visual inspection.

Discussion

Summary

This re-identification uses de-identified hospital discharge data and a simulated HIPAA version of the data. First, we demonstrate that de-identification protocols do not provide reasonable expectations of privacy protections for patients even though the datasets may contain redacted or aggregated data. Second, we recognize that states tend to follow variations that are close to the HIPAA Safe Harbor guidelines, and so create HIPAA versions of the datasets. We re-identified 28.3 percent of named individuals from Maine news article and 34 percent of the named individuals from Vermont. When using HIPAA versions of the data, we re-identified 3.2 and 10.6 percent of named individuals in Maine and Vermont, respectively.

Maine and Vermont's re-identification rates are lower than the Washington State study, which re-identified 43 percent of named individuals. This difference may be due to the fact that Washington's data reported age using month and year. Maine shared patient age and

Vermont shared patient age using age groups. However, state data using standards similar to HIPAA still are not immune from re-identification. Our results demonstrate that even states that generalized demographic data such as age can have high re-identification rates. So, the methods of this study show that we can uniquely re-identify patients even when these state data use HIPAA Safe Harbor standards.

Our results demonstrate that HIPAA Safe Harbor is not the guideline or framework for effectively anonymizing data to prevent re-identifications. What may be needed is the flexibility to consider the rapidly increasing availability of data when assessing de-identification methods instead of a fixed standard of de-identification checklists, as HIPAA Safe Harbor suggests. While we used the HIPAA Safe Harbor for our test, we do not suggest that the HIPAA Safe Harbor is the ideal standard.

Further Risks to Privacy

Re-identified hospital records for motor vehicle or home accidents sometimes included diagnosis codes describing pre-existing conditions that have nothing to do with the patient's reason for hospitalization, such as those detailing alcoholism, diabetes, and depression. For instance, after uniquely re-identifying a hospital record concerning a pedestrian who was hit by a car, we learned that the person involved in the car crash was a college student with a previous history of depression, PTSD, and anxiety (see Figure 24). Such associations can be damaging, especially when re-identified hospital records reveal sensitive or stigmatized mental health histories. In one instance, a political candidate's campaign was marred when an attacker exposed her prior suicide attempt [28]. Even though this political candidate won in the end, and some may argue that this incident helped her campaign, her mental health was exposed without her consent. Diagnoses codes can also include information about unemployment status, cannabis and opioid use, and histories of physical or domestic violence.

Maine	NewsData	HospitalData
Demographics	Name	--
	19-year-old	Age: 19
	Female	Gender: F
	Bangor, Maine	Geocode: 19020
Hospital Information	York Hospital	HP: 200020
E-code	"hit by a car while crossing highway"	E8187 Other noncollision motor vehicle traffic accident injuring pedestrian
Diagnoses	"two broken bones in right leg, bruises to arms, legs, suffered head injuries"	87344 Open wound of jaw, without mention of complication
		82380 Closed fracture of unspecified part of tibia alone
		8249 Unspecified fracture of ankle, open
		88101 Open wound of elbow, without mention of complication
Sensitive information unrelated to hospitalization included in HospitalData		311 Depressive disorder
		30981 Posttraumatic stress disorder
		30000 Anxiety state

Figure 24. Example re-identification that revealed sensitive information about a patient involved in a motor vehicle accident.

Re-identification may also become easier as hospitals use more specific diagnoses and procedure codes. The new edition of the International Classification of Diseases, 10th edition (or ICD10) [29], is more detailed than the ICD9 codes that we used in this re-identification strategy. The ICD10 codes allow hospitals to indicate which leg (left or right) sustained injuries. ICD9 codes did not allow for that level of distinction. It may in fact be faster to verify diagnoses codes using more specific ICD10 codes.

Newspaper articles are a unique way to re-identify data because they routinely report and publish detailed accounts of accidents that result in a hospital visit. This re-identification strategy relies on the quality and specificity of the news stories. We found that more specific injury details from the news stories rendered more diagnoses fields that we could use to confirm matches that were based on age, gender, geocode or ZIP code, admission quarter and hospital. The solution is not to censor journalists from reporting local events nor is it to stop data sharing altogether. Our findings suggest that patients are vulnerable to re-identification not just due to the specific strategy in this paper which uses news stories, but also to other sources of the same kind of information about hospitalizations and medical conditions, including prescription data, medical mailing lists, employers, debtors, friends, and family.

The Role of Re-identifications

Currently, some organizations have legally prohibited re-identification of state records because they are disruptive. Maine has the following clause in their data-use agreement [30]:

> 2. DATA PRIVACY AND DATA SUPPRESSION
>
> C. The Data Applicant and Data Recipient will not use the MHDO Data in any way, or link these data to other records or data bases, if the result allows for identifying individuals, unless authorized in writing by the MHDO.

States can reduce disruption, not by banning re-identification experiments, but by encouraging researchers to report vulnerabilities. More sensitive data should be made available under an application process that asks for research or use purposes. A data-use agreement should allow recipients who apply to run approved re-identification experiments on the condition that they report results to the state and that they not disclose the re-identified patients. There can be penalties ranging from a loss of access to the data to cease-and-desist orders prohibiting the data from being used if re-identification results are not reported, or the data is not used for stated purposes.

Such an agreement can also direct anyone holding a vulnerable version of the data to stop using it once a vulnerability is confirmed until a better-redacted version is provided or they qualify for an exemption. Doing so encourages responsible sharing in a manner that allows data holders to improve with scientific knowledge and innovation. States and legislatively-mandated agencies or organizations can also do these tests internally as well, since they can access full datasets and assess different levels of data suppression. They can determine how much information to distribute to maximize both the utility of the data and the privacy of the patients.

Vermont does not require data-use agreements or have data access restrictions. Vermont's existing protocols, while not necessarily compliant with HIPAA Safe Harbor, do apply simple aggregations (from ZIP code to "ZIP code group" and age to "Age Group) and redactions (no personal identifiers such as name or address included). Re-identification experiments such as this study provide the evidence needed to improve data sharing practices.

One state that reviewed its de-identification practices since 2012 is Maine [31]. In 2016, Maine implemented new rules establishing different levels of statewide hospital datasets available for public access (see Appendix C for all data fields included in the three data levels), requiring applications and data-use agreements, publishing a list of data requesters and recipients, and allowing a comment period so data providers and the public could to voice concerns to the MHDO about the release of data to anyone requesting it. These new rules are an effective improvement. Designating data levels and creating more stringent requirements

to acquire sensitive data may better deter re-identification of patient data. Tiered-access for data sharing has the potential to be effective in communicating clear access requirements [32]. Data-purchase applications and data-use agreements also establish legal protections and inform recipients of the expectations for privacy when using the data. Publishing a list of data requestors and recipients permits the public to participate in holding both the MHDO and data requestors accountable and transparent about purposes and uses of the data.

Despite these new rules, Maine's lowest tiered data access still does not properly conform to HIPAA Safe Harbor because it contains admission and discharge quarter and year, and health service area. There may be utility in receiving dates in quarters or at other levels more granular than years to researchers looking at occurrence rates of seasonal diseases, but this paper suggests that including admission quarters increases the identifiability of a dataset. Our study found 26 unique re-identified matches in the Maine data and eight unique re-identified matches in the simulated HIPAA data that did not include admission quarter. Our results using the Maine HIPAA Dataset, suggest that our re-identification strategy would still be successful using the most unrestricted version of the Maine dataset.

The costs for unrestricted hospital data are low. Sweeney obtained the data for Washington State for only $50 and Maine's 2010 hospital data was $1,125. Vermont's data for Inpatient, Outpatient, Emergency Department, Expanded Outpatient and Revenue Data for 2006 through 2014 is all available online for free [23]. A 2012 survey of inpatient unrestricted data costs for 33 states reports that the total cost to obtain data from all 33 states would be between $39,729 and $48,705 [2]. The costs ranged from $0 (Vermont and New Hampshire) to $10,000 (Tennessee).

Beyond statewide health data, many academic publications ask authors to submit a version of the data used to report result as a condition of publication (for examples, see [33]) to verify published findings, promote further studies, and possibly lower research costs by allowing data reuse. In the recent years, government offices at all levels, enforcement agencies, and non-profit organizations are launching open data portals for the sake of transparency, accountability, and research.

HIPAA Safe Harbor's framework is often considered the *de facto* standard for protecting patient privacy even though HIPAA has not been rigorously confirmed to guarantee privacy [34][35]. However, these standards have not been rigorously tested. There are scientifically tested privacy solutions that do guarantee privacy protections, such as k-anonymity, which uses either generalization or suppression of fields to guarantee a defined level of anonymity to individual records in a given dataset [36]. Differential privacy is another method using synthetic datasets that still retain statistical properties of the original dataset [37]. All data-sharing methods should continuously undergo a cycle of testing and subsequent improvement in order to guarantee formal privacy protections like these.

A re-identification experiment helped Washington State to improve its data sharing practices. We hope these experiments will similarly help Maine and Vermont to elect an improved privacy approach. At large, our results suggest that all states that share hospital data revisit vulnerabilities and de-identification practices. The de-identification checklist that HIPAA Safe Harbor promotes is the bare minimum protection against re-identification. Policy-makers and data-sharing centers should consider scientifically tested protocols that guarantee privacy protections to patients, especially since they cannot opt out of inclusion in hospital records.

References

1. Love D, Sullivan E, Davis R, Prysunka A, and Rudolph B. The Collection and Release Practices of Physician Identifiers in Statewide Hospital Discharge Data Reporting. 2010. https://www.nahdo.org/sites/nahdo.org/files/Resources/Data_Analytics/The%20Collection%20and%20Release%20Practices%20onf%20MDID%20in%20HDD.pdf

2. Hooley S and Sweeney L. Survey of Publicly Available State Health Databases. 2010. dataprivacylab.org/projects/50states

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

4. Andrews R. Statewide Hospital Discharge Data: Collection, Use, Limitations, and Improvements. HSR: Health Services Research 50:S1, Part II (August 2015). DOI: 10.1111/1475-6773.12343

5. Basu J and 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.

6. 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.

7. 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.

8. 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.

9. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

10. Robertson J. Public Records Requests for State Discharge Data (updated with Maine). Bloomberg News. ForeverData.org. Collection 1007. November 2012.

11. Robertson J. Who's Buying Your Medical Records. Bloomberg News. June 5, 2013. https://www.bloomberg.com/news/articles/2013-08-08/your-medical-records-are-for-sale

12. Robertson J. States Hospital Data for Sale Puts Privacy in Jeopardy. Bloomberg News. Also Bloomberg BusinessWeek. June 5, 2013. https://www.bloomberg.com/news/articles/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy

13. comprehensive Hospital Abstract Reporting System: Hospital Inpatient Dataset: Clinical Data. Washington State Department of Health. http://www.doh.wa.gov/DataandStatisticalReports/HealthcareinWashington/HospitalandPatientData/HospitalDischargeDataCHARS

14. Public Use File Request. California Office of Statewide Health Planning and Development. 2016. https://www.oshpd.ca.gov/HID/Data_Request_Center/PUF.html

15. Data Availability. Maine Health Data Organization. 2013. https://mhdo.maine.gov/tableau/data.cshtml

16. Pricing Information. Maine Health Data Organization. 2013. https://mhdo.maine.gov/pricing_information.html

17. 2010 Unrestricted Hospital Inpatient Data Release. Maine Health Data Organization. 2013. https://mhdo.maine.gov/inpatient_outpatient.htm

18. Data Elements Specifications Workbook. Maine Health Data Organization. 2013. https://mhdo.maine.gov/_dataRqstDocs/ReleaseIncludedElements_161013.xlsx

19. Claim Inpatient Admission Type Code. Research Data Assistance Center, Center for Medicaid and Medicare Center. https://www.resdac.org/cms-data/variables/Claim-Inpatient-Admission-Type-Code

20. Hospital 2012 Inpatient Refresh Release Notes. Maine Health Data Organization. 2013. https://mhdo.maine.gov/inpatient_outpatient.htm

21. Maine Office of GIS Open Data. Office of Information Technology. 2017. https://geolibrary-maine.opendata.arcgis.com/datasets/dda746a3f11a4fde8067e01507a0c879_4

22. Population, Housing Units, Area, and Density: 2010 - State -- Place and (in selected states) County Subdivision. 2010 Census Summary File 1. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk

23. Vermont Uniform Hospital Discharge Data Sets. Vermont Department of Health. 2017. http://www.healthvermont.gov/health-statistics-vital-records/health-care-systems-reporting/hospital-discharge-data

24. ICD-9-CM Diagnosis and Procedure Codes: Abbreviated and Full Code Titles. Center for Medicaid and Medicare Center. https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes.html

25. LexisNexis. http://www.lexisnexis.com/hottopics/lnacademic/

26. Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000. http://dataprivacylab.org/projects/identifiability/

27. Sweeney L. Patient Identifiability in Pharmaceutical Marketing Data. Data Privacy Lab Working Paper 1015. Cambridge 2011. https://dataprivacylab.org/projects/identifiability/pharma1.html

28. Newman M. Candidate Faces Issue of Suicide. October 10, 1992. http://www.nytimes.com/1992/10/10/nyregion/candidate-faces-issue-of-suicide.html?mcubz=3

29. Preparing for the ICD-10 Code Set: October 1, 2014 Compliance Date. American Medical Association. 2012. https://www.unitypoint.org/waterloo/filesimages/for%20providers/icd9-icd10-differences.pdf

30. Request Data or Reports. Maine Health Data Organization. 2013. https://mhdo.maine.gov/data_rqst_process.htm

31. Rule Chapter 120 - Release of Data to the Public. Maine Health Data Organization. 2013. https://mhdo.maine.gov/_finalStatutesRules/Chapter%20120%20Data%20Release_Effective160728%20.docx

32. Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The DataTags System. Technology Science. 2015101601. October 16, 2015. http://techscience.org/a/2015101601

33. PLOS PLoS One. Data Availability. http://journals.plos.org/plosone/s/data-availability

34. Garfinkel S. NISTIR 8053 De-Identification of Personal Information. National Institute of Standards and Technology, U.S. Department of Commerce. http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

35. Honigman Miller Schwartz and Cohn's HIPAA LAW FOCUS. A Special Joint Newsletter on the HIPAA Privacy Rule Prepared by the Health Care and Employee Benefits Departments of HMS. August 2002. https://www.honigman.com/media/site_files/396_imgimgHIPAA_20Law_20Focus_20-_20August_202002.pdf

36. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588. http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity2.html

37. Dwork C. Differential Privacy: a survey of results. International Conference on Theory and Applications of Models of Computation TAMC 2008: pp 1-19.

Authors

Ji Su Yoo is an experienced Research Analyst in privacy, security, and technology at Harvard University's Data Privacy Lab and at the Institute for Quantitative Social Science. She graduated cum laude from Harvard College with an A.B. in Social Studies. She has written papers and given talks on the intersection between government and technology, especially in relation to privacy and data analysis in health care. Currently, she designs research projects to expose the unintended consequences of technology and how bias can exacerbate and maintain existing political and societal inequalities. Ji Su wants to continue conducting research that will inform public interest issues. As managing editor of *Technology Science*, Ji Su Yoo was recused from the review of this paper.

Alexandra Thaler is a junior at Harvard College. She is pursuing a concentration in government with a secondary in global health and health policy. She has worked in startups doing international market research and comparative pricing analysis, and has also worked as a research assistant in a neurobiology lab at the University of Pennsylvania Medical School, focusing on cued fear in animal models of concussion. While taking Professor Latanya

Sweeney's course "The Politics of Personal Data," she became excited about the intersection of her interests in health, tech, and policy, which led to her project on re-identification of health care data. At Harvard, Alexandra played on the varsity lacrosse team and worked for the Harvard Alumni Association.

Latanya Sweeney is professor of government and technology in residence at Harvard University, X.D. and Nancy Yang Faculty Dean of Currier House at Harvard, director of the Data Privacy Lab at Harvard, and editor in chief of *Technology Science*. She was formerly the chief technology officer of the U.S. Federal Trade Commission. She earned her Ph.D. 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. As editor in chief of *Technology Science*, Professor Sweeney was recused from the review of this paper.

Jinyan Zang is a PhD student in Government at Harvard University and a 2017 recipient of the Paul and Daisy Soros Fellowships for New Americans. He is an experienced researcher on consumer protection, data security, and privacy issues as a Research Fellow at the Federal Trade Commission and a Research Analyst at Harvard University's Data Privacy Lab and at the Institute for Quantitative Social Science. He graduated cum laude in 2013 from Harvard College with a B.A. in Economics. As Managing Editor of *Technology Science*, Jinyan Zang was recused from the review of this paper.

The initial iteration of data entry and matching was performed by Robert Collins, Louis Delano, Eugenio Donati, Becina Ganther, Harshita Gupta, Carl Sibley, and Hirsh Sisodia, as students who participated in one of the projects in Harvard College Freshman Seminar course 42R, Technology to Save the World, in the Spring semester of 2017.

Citation

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 09, 2018. http://techscience.org/a/2018100901

Appendix

A. 2010 Maine Inpatient Data Fields

Field Number	Field Name	Field Description	Example Value
1	mhdoid	MHDO Physical Record Number	110240020795
2	hp	Hospital Code	200024
3	sx	Patient Sex	F
4	tadm	Priority of Visit (Type)	2
5	sadm	Point of Origin of Admission (Source)	4
6	adx	Admitting Diagnosis	80131
7	xref35	Hospital Service Area of Patient	3
8	xref66	Health Planning Area of Patient	23
9	geocode	Patient Geo Code	1050
10	disp	Patient Disposition/Discharge Status	62
11	pay1	Principal Payer Code Aggregated	1
12	naic1	Principal Payer NAIC number	180
13	pay2	Secondary Payer Code Aggregated	1
14	naic2	Secondary Payer NAIC number	180
15	pay3	Tertiary Payer Code Aggregated	6
16	naic3	Tertiary Payer NAIC number	COM00
17	dx1	Principal or First Diagnosis submitted	80131
18	poa1	Principal Diagnosis Present on Admission	Y
19	dx2	Second Diagnosis submitted	E8809
20	poa2	Second Diagnosis POA Code	Y
21	dx3	Third Diagnosis submitted	E8490
22	poa3	Third Diagnosis POA Code	1
23	dx4	Fourth Diagnosis submitted	5990
24	poa4	Fourth Diagnosis POA Code	N
25	dx5	Fifth Diagnosis submitted	7804
26	poa5	Fifth Diagnosis POA Code	Y
27	dx6	Sixth Diagnosis submitted	2768
28	poa6	Sixth Diagnosis POA Code	N

Field Number	Field Name	Field Description	Example Value
29	dx7	Seventh Diagnosis submitted	4019
30	poa7	Seventh Diagnosis POA Code	Y
31	dx8	Eighth Diagnosis submitted	30000
32	poa8	Eighth Diagnosis POA Code	Y
33	dx9	Ninth Diagnosis submitted	3899
34	poa9	Ninth Diagnosis POA Code	Y
35	dx10	Tenth Diagnosis submitted	
36	poa10	Tenth Diagnosis POA Code	
37	dx11	Eleventh Diagnosis submitted	
38	poa11	Eleventh Diagnosis POA Code	
39	ps1	Principal Procedure Code	
40	ps2	Second Procedure Code	
41	ps3	Third Procedure Code	
42	ps4	Fourth Procedure Code	
43	ps5	Fifth Procedure Code	
44	ps6	Sixth Procedure Code	
45	pd1	Principal Procedure Code Date	
46	pd2	Second Procedure Code Date	
47	pd3	Third Procedure Code Date	
48	pd4	Fourth Procedure Code Date	
49	pd5	Fifth Procedure Code Date	
50	pd6	Sixth Procedure Code Date	
51	apspec	Attending Physicians Specialty	60
52	spspec	Surgeon/Other Provider Specialty	
53	arev1	Accommodations Rev Code - 1	120
54	arev2	Accommodations Rev Code - 2	200
55	arev3	Accommodations Revenue Code - 3	206
56	days1	Accommodations Days -1	1

Field Number	Field Name	Field Description	Example Value
57	days2	Accommodations Days -2	1
58	days3	Accommodations Days -3	1
59	apdrg	All-Patient DRG	766
60	apmdc	All-Patient MDC	1
61	los	Length of Stay	3
62	age	Patients Age	77
63	wgt	Estimated Birth weight < 30 days old	
64	race	Patient's Race (coded)	5
65	ethn	Patients Ethnicity (coded)	2
66	msdrg	CMS DRG	86
67	msmdc	CMS MDC	1
68	pname1	Name of the Principal Payer	MEDICARE PART A
69	rvhits	Total Number of Ancillary Revenue Codes	18
70	rv1	Ancillary Revenue Codes	250
71	rv2	Ancillary Revenue Codes	258
72	rv3	Ancillary Revenue Codes	259
73	rv4	Ancillary Revenue Codes	270
74	rv5	Ancillary Revenue Codes	300
75	rv6	Ancillary Revenue Codes	301
76	rv7	Ancillary Revenue Codes	305
77	rv8	Ancillary Revenue Codes	306
78	rv9	Ancillary Revenue Codes	307
79	rv10	Ancillary Revenue Codes	320
80	rv11	Ancillary Revenue Codes	350
81	rv12	Ancillary Revenue Codes	420
82	rv13	Ancillary Revenue Codes	424
83	rv14	Ancillary Revenue Codes	430
84	rv15	Ancillary Revenue Codes	636

Field Number	Field Name	Field Description	Example Value
85	rv16	Ancillary Revenue Codes	730
86	rv17	Ancillary Revenue Codes	762
87	rv18	Ancillary Revenue Codes	960
88	rv19	Ancillary Revenue Codes	
89	rv20	Ancillary Revenue Codes	
90	rv21	Ancillary Revenue Codes	
91	rv22	Ancillary Revenue Codes	
92	rv23	Ancillary Revenue Codes	
93	rv24	Ancillary Revenue Codes	
94	rv25	Ancillary Revenue Codes	
95	rv26	Ancillary Revenue Codes	
96	rv27	Ancillary Revenue Codes	
97	rv28	Ancillary Revenue Codes	
98	rv29	Ancillary Revenue Codes	
99	rv30	Ancillary Revenue Codes	

A. 2011 Vermont Emergency Department Data Fields

Field Number	Field Name	Field Description	Example Value
1	HNUM2	HOSPITAL_NUMBER	8
2	ATYPE	ADMISSION_TYPE	1
3	ASOUR	ADMISSION_SOURCE	3
4	AGEGRP	Age Group	9
5	ZIPGRP	Most ZIP code groups are the first 3 digits of the ZIP code; ZIP codes between 05800 and 05999 are grouped into "058-059"	57
6	SEX	SEX	1
7	DSTAT	DISCHARGE_STATUS	5
8	PPAY	PRINCIPAL_PAYMENT_SOURCE	6

Field Number	Field Name	Field Description	Example Value
9	CHRGS	TOTAL_CHARGES (adjusted)	14580
10	SCUDA	NUMBER OF DAYS IN SPECIAL CARE UNIT, Includes all CCU, ICU & other SCU (INPATIENT ONLY)	843
11	DRG	DIAGNOSIS RELATED GROUP (INPATIENT ONLY)	5566
12	MDC	MAJOR DIAGNOSTIC CATEGORY (INPATIENT ONLY)	5990
13	DX1	Principal Diagnosis	2761
14	DX2	2nd Diagnosis	2768
15	DX3	3rd Diagnosis	V5861
16	DX4	4th Diagnosis	7850
17	DX5	5th Diagnosis	V433
18	DX6	6th Diagnosis	2724
19	DX7	7th Diagnosis	4019
20	DX8	8th Diagnosis	60784
21	DX9	9th Diagnosis	
22	DX10	10th Diagnosis	
23	DX11	11th Diagnosis	
24	DX12	12th Diagnosis	
25	DX13	13th Diagnosis	
26	DX14	14th Diagnosis	
27	DX15	15th Diagnosis	
28	DX16	16th Diagnosis	
29	DX17	17th Diagnosis	
30	DX18	18th Diagnosis	
31	DX19	19th Diagnosis	
32	DX20	20th Diagnosis	
33	PX1	Principal Procedure	
34	PX2	2nd Procedure	
35	PX3	3rd Procedure	
36	PX4	4th Procedure	

Field Number	Field Name	Field Description	Example Value
37	PX5	5th Procedure	
38	PX6	6th Procedure	
39	PX7	7th Procedure	
40	PX8	8th Procedure	
41	PX9	9th Procedure	
42	PX10	10th Procedure	
43	PX11	11th Procedure	
44	PX12	12th Procedure	
45	PX13	13th Procedure	
46	PX14	14th Procedure	
47	PX15	15th Procedure	
48	PX16	16th Procedure	
49	PX17	17th Procedure	
50	PX18	18th Procedure	
51	PX19	19th Procedure	
52	PX20	20th Procedure	
53	ECODE or ECODE1	E-code1	9
54	ECODE2	E-code2, added in 2010	5
55	ECODE3	E-code3, added in 2010	135
56	HSA	HOSPITAL_SERVICE_AREA or HSA4..	9
57	PDAYS	(DISCHARGE DATE minus ADMISSION DATE) 1 if discharged on date of admission or on day after admission Length of stay greater than 255 days was recoded to 255.	
58	SDF	SAME DAY FLAG (INPATIENT ONLY)	
59	GROUPER	Version Number corresponding with discharges within a date range (INPATIENT ONLY)	
60	CCSDX	CCS DIAGNOSIS (DX) GROUP http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp	
61	CCSDXGRP	CCS DX HIGH LEVEL GROUP	2011

Field Number	Field Name	Field Description	Example Value
62	CCSPX	CCS PROCEDURE (FIRST-LISTED PX) GROUP	36675
63	CCSPXGRP	CCS PX HIGH LEVEL GROUP	111
64	CCSPPX	CCS PROCEDURE (PPX - FIRST PX IN RANGE 00.00-86.99) GROUP	1
65	CCSPPXGRP	CCS PPX HIGH LEVEL GROUP	0
66	CCSPROC	NEW IN 2014 DUE TO CHANGES IN REPORTING OF OUTPATIENT PROCEDURES	1
67	CCSPROCGRP	NEW IN 2014 DUE TO CHANGES IN REPORTING OF OUTPATIENT PROCEDURES	0
68	DY	For 2000 and later, full data year (1999 and prior was 2 digits)	0
69	RECNO	RECORD NUMBER	36675
70	BTYPE	Bill type as designated by the hospital	2
71	ERFLAG	Had an associated Emergency Room Code (45x)	2
72	CAH	Critical Access Hospital	14580
73	STRES	State of Residence	
74	OBSFLAG	Had an associated Observation Room code (760 or 762)	
75	AFLAG	ambulatory surgery' flag	
76	UNIQ	Unique Record Number, for linking to revenue file	
77	ADMID_QTR	Admission Quarter	
78	DISCD_QTR	Discharge Quarter	
79	CHRGS_HCIA	Unadjusted Total Charges	

A. 2016 Maine MDHO Data Elements Released in Levels I, II, and III

Hospital Inpatient Level I Fields

Level	Data Element	Element Long Name	Notes
Level I	sysid	Unique Record Number	
Level I	mhdoid	MHDO Physical Record Number	
Level I	hp	Hospital Code	

Hospital Inpatient Level I Fields

Level	Data Element	Element Long Name	Notes
Level I	sx	Gender	
Level I	age	Age	
Level I	tadm	Priority of Visit (Type)	
Level I	sadm	Point of Origin of Admission (Source)	
Level I	adx	Admitting Diagnosis	
Level I	admyear	Admission Year	Only received if not getting Level II Data
Level I	admqtr	Admission Quarter	Only received if not getting Level II Data
Level I	disyear	Discharge Year	Only received if not getting Level II Data
Level I	disqtr	Discharge Quarter	Only received if not getting Level II Data
Level I	disp	Disposition/Discharge Status	
Level I	pay1	Principal Payer Code Aggregated (MHDO code)	
Level I	pay2	Secondary Payer Code Aggregated (MHDO code)	
Level I	pay3	Tiertiary Payer Code Aggregated (MHDO code)	
Level I	ds1	Principal or First Diagnosis Submitted	
Level I	poa1	Principal Diagnosis Present on Admission (POA) Code	
Level I	ds2-ds9	Additional Diagnoses Submitted	
Level I	poa2-poa9	Additional Diagnoses POA Code(s)	
Level I	ecode1-ecode2	External Cause Code(s)	
Level I	poae1-poae2	External Cause POA Code(s)	
Level I	ps1	Principal Procedure Code	
Level I	ps2-ps6	Additional Procedure Code(s)	
Level I	apspec	Attending Physicians Specialty	
Level I	spspec	Surgeon/Other Provider Specialty	

Hospital Inpatient Level I Fields

Level	Data Element	Element Long Name	Notes
Level I	arev1-arev3	Accommodations Revenue Code(s)	
Level I	days1-days3	Accommodations Day(s)	
Level I	apdrg	DRG - All Patient	
Level I	apmdc	MDC - All Patient	
Level I	los	Length of Stay	
Level I	wgt	Estimated Birth Weight (< 30 Days Old)	
Level I	msdrg	MS-DRG	
Level I	mdmdc	MS-MDC	
Level I	rvhits	Total Number of Ancillary Revenue Code(s)	
Level I	rv1-rv50	Ancillary Revenue Code(s)	
Level I	cms-drg	CMS DRG	
Level I	cms-mdc	CMS MDC	
Level I	apr-drg	APR-DRG	
Level I	apr-mdc	APR-MDC	
Level I	tax01	Taxonomy code for attending provider	
Level I	tax02	Taxonomy code for secondary provider	

Hospital Inpatient Level II Fields

Level	Data Element	Element Long Name	
Level II	idnox	Patient Medical Record Number (Encrypted)	
Level II	dob	Date of Birth	Request for this field requires additional explanation
Level II	county	County of Patient	
Level II	geocode	Maine GIS geocode of Patient	Request for this field requires additional explanation

Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: a re-identification of patients in Maine and Vermont statewide hospital data. Technology Science. 2018100901. October 9, 2018. http://techscience.org/a/2018100901

Hospital Inpatient Level II Fields

Level	Data Element	Element Long Name	
Level II	state	Patient State	Request for this field requires additional explanation
Level II	city	Patient City	Request for this field requires additional explanation
Level II	ZIP	US Postal code of Patient	Request for this field requires additional explanation
Level II	indate	Admission Date	
Level II	inhour	Admission Hour	
Level II	otdate	Discharge Date	
Level II	othour	Discharge Hour	
Level II	pd1	Principal Procedure Code Date	
Level II	pd2-pd6	Additional Procedure Code Date(s)	
Level II	dcode1	Attending Physician Code (Encrypted)	
Level II	dcode2	Servicing Physician Code (Encrypted)	
Level I	xref35	Hospital Service Area of Patient	Defined as Level I in Chapter 120, but released w. Level II
Level I	xref66	Health Planning Area of Patient	Defined as Level I in Chapter 120, but released w. Level II

Hospital Inpatient Level III Fields

Level	Data Element	Element Long Name
Level III	TBD	Self-Pay Name (individuals who are payers in hospital data)
Level III	TBD	Race (coded)
Level III	TBD	Ethnicity (coded)