Patient Identifiability in Pharmaceutical Marketing Data

Latanya Sweeney, PhD

Carnegie Mellon University
Harvard University

Abstract

Does pharmaceutical marketing data expose patient records? In 2003, just after the promulgation of the HIPAA Privacy Rule, a major American pharmaceutical company commissioned a report across 9 states to determine the number of people in those states who may be at risk of being identified if patient pharmacy claims data used for marketing were shared. In May 2003 the report showed that 2.3% of individuals could be uniquely identified from the de-identified prescription records used for marketing purposes at the time and that 6.1% were identifiable to a binsize of 2 (i.e., the record either uniquely related to one named person or related indistinguishably to 2 identified people). These results used prescription information {*drug*, *dosage and refill information*, *patient diagnosis*, patient ZIP inferred from *pharmacy ZIP*, *prescription fill date*}. No explicit patient identifiers (e.g., name or address) appeared in the data. The prescribing doctor was not uniquely identified. Results were based on the states: New York, Illinois, Michigan, Massachusetts, Florida, California, Pennsylvania, Texas, and Arizona. The primary means of re-identification was linking the prescription records to ambulatory and hospital discharge data using patient {*diagnosis*, *inferred ZIP*, and *drug*, *dosage and refill information*} to learn more patient demographics and then linking that result to a voter list (or other population register) to learn the names of the subjects of the prescriptions. In comparison, the HIPAA Safe Harbor tends to re-identify about 0.04% of the population, thereby showing that in general more personal information is put at risk in these data than with the HIPAA Safe Harbor, however variability exists in re-identification rates from state to state with some states having re-identification rates less than the HIPAA Safe Harbor. Other privacy observations found in the data, but not part of the analysis, include: (1) the data did not segment or restrict access to special medical classes protected by law, such as psychiatric and HIV related prescriptions; and, (2) the data made it possible to construct a patient's prescription profile over time, which could further increase re-identification risk. This paper summarizes the earlier 2003 report, reviews subsequent publication, and imposes the emergent scientific-legal approach of comparing re-identification rates to the HIPAA Safe Harbor. In the end though, this paper demonstrates the best of measuring de-identification risks while exposing the perils of de-identification as a regime.

Keywords: HIPAA Privacy Rule, identifiability, data privacy, re-identification

1. Introduction

Price Waterhouse Coopers predicts that sharing personal health information beyond the direct care of the patient will be a two billion dollar market over the next few years [1]. Many companies thrive through selling data based on acquiring, curating and aggregating personal data. For example, IMS Health collects personal prescription information from pharmacies and pharmacy benefits programs, and then uses it to sell market information to pharmaceutical companies [2]. Acxiom collects personal information from public records, such as marriage licenses and voter lists, and uses it to provide background checks [3]. Geisinger Health System, a large integrated health system, created a company called MedMining, which licenses its data to promote healthcare research, primarily to major pharmaceutical companies and large biotech companies [4].

A key question is whether shared data in today's data affluent environment respects the intent of privacy regulations. For years, privacy policies relied on de-identification, the removal of explicit identifiers (e.g. name, address, and Social Security number), as a way to provide privacy in data. This approach is too naive in today's data rich society because other data sources often exist that contain some or all of the same values, allowing redacted identity information to be restored by linking datasets. As evidence, there have been several highly publicized cases of re-identifications [e.g. 5, 6, 7, 8]. Stronger privacy technology protections based, such as k-anonymity [9] and differential privacy [10], modify the data beyond merely removing explicit identifiers but incentives are lacking to use or develop these technical approaches as access to poorly de-identified data remains widespread.

For insight, we revisit prior work done in 2003, just after the Health Insurance Portability and Accountability Act Privacy Rule ("HIPAA") was promulgated [11], and determine how the identifiability of pharmaceutical marketing data based on de-identified patient prescription data compares with acceptable levels of identifiability derived from HIPAA.

2. Background

In 2003, a major pharmaceutical company commissioned a report across 9 states to determine the number of people in those states who may be at risk of being identified if patient pharmacy claims data used for marketing were shared. The pharmaceutical company had contracts with a number of managed care organizations in which the managed care organizations received rebates that depended on their use of the pharmaceutical company's products. These contracts required the managed care organizations to submit copies of patient prescription claims data after each quarter to allow the pharmaceutical company to validate and ultimately pay rebates as established by the contract.

The pharmaceutical company was concerned about sharing the data under HIPAA and asked researchers at two different organizations for a two-tiered determination as to whether values in the data were sufficiently de-identified under HIPAA so that the data

could be shared and used broadly. One group, led by Dr. Sweeney [12], computed identifiability risks, and the other group, led by Dr. Stoto [13], mapped those risks to the statistical disclosure literature at the time [14]. This report summarizes the efforts of both groups and updates the results based on more recent approaches [15][16][17][18][19], all of which is further discussed in the later sections of this writing after describing the dataset and legal standards below.

2.1. Dataset Fields

The data, termed the "Dataset", consists of a record for each prescription filled. Figure 2 provides the overall domain of fields available and Figure 1 shows the 19 fields that serve as the basis for the assessment. Presumably the pharmaceutical company could achieve its marketing objectives using the fields in Figure 1 only.

Fields in Figure 2 and Figure 1 report information about the patient, the prescribing physician, and the pharmacy. No explicit patient identifiers, such as name and address, appear, but there is patient specific information in the overall layout (Figure 2). Specifically, *diagnosis* and *prescription_number* and possibly *contract_id* and *group_id*, because *contract_id* may identify the patient's employer and *group_id* may be the number assigned to the patient's family. In the fields that are the subject of the assessment (Figure 1), *diagnosis* is the only explicit patient information. The prescribing physician is explicitly identified in the overall layout (Figure 2) by *prescriber_id*, which is a commonly used industry number that appears in a publicly available registry of explicitly identified physicians, but that information does not appear in Figure 1. The pharmacy that filled the prescription is explicitly identified in Figure 2 by *pharmacy_id* , which is a commonly used industry number that appears in a publicly available registry of explicitly identified pharmacies. Only the field *pharmacy_ZIP* appears in Figure 1. All other fields in Figure 1 refer to drug information. In summary, the subject of the assessment, unless stated otherwise or made obvious from context, uses the fields in Figure 1, which includes the patient's diagnosis and prescription number, the pharmacies explicit identity and ZIP, drug and refill information, and the date (day, month and year) the prescription was filled.

1. Prescription number
2. Pharmacy ID
3. Date of fill
4. NDC number
5. Quantity
6. Plan/Prescription level
7. Plan ID
8. Plan Name
9. Pharmacy ZIP
10. Unit of measure
11. Dosage form
12. Diagnosis code
13. Days supply
14. Prescription type
15. Total number of prescriptions
16. Therapeutic class
17. Reimbursement date
18. New/refill code
19. Product description

Figure 1. Subset of fields described in Figure 2 that are the subject of the assessment.

Field Number	Field	field name	Start Position	Length	End Position	Contents	Format
1	RECORD TYPE	rectype	1	2	2	'UD'	
2	LINE NUMBER	linenum	3	11	13	SPACES: unique claim id resides in cols 382-400	
3	DATA LEVEL	level	14	2	15	'PP'	
4	PLAN ID QUALIFIER	planidq	16	1	16	'C' - CONTRACTING	
5	PLAN ID CODE	planidc	17	17	33	CARRIER_OPERATIONAL_ID	
6	PLAN NAME	plan	34	30	63	CARRIER_NME	
7	PHARMACY ID QUALIFIER	pharmidq	64	1	64	'Z'	
8	PHARMACY ID CODE	pharmidc	65	17	81	NABP_NBR	
9	PHARMACY ZIP CODE	pharmzip	82	9	90	CLAIM_POSTAL_CDE	
10	PRODUCT CODE QUALIFIER	ndcq	91	1	91	'N'	
11	PRODUCT CODE	ndc	92	17	108	FILL_NDC_NBR	
12	PRODUCT DESCRIPTION	desc	109	30	138	LABEL_TXT	
13	DAW PRODUCT SELECTION CODE	daw	139	1	139	BILLING_DAW_CDE	
14	TOTAL QUANTITY	qty	140	15	154	INFERRED_FILL_QTY	9(11)V999-
15	UNIT OF MEASURE	unit	155	2	156	'EA'	
16	DOSAGE FORM ID CODE	dose	157	2	158	DOSAGE_FORM_CDE	
17	DIAGNOSIS CODE	dx	159	6	164	SPACES	
18	REBATE DAYS SUPPLY	rebatday	165	4	168	FILL_DAYS_SUPPLY_QTY	9(3)-
19	PRESCRIPTION TYPE	prestyp	169	2	170	SPACES	
20	TOTAL NUMBER OF PRESCRIPTIONS	totnum	171	8	178	CLAIM_COUNT_NBR	9(7)-
21	PRESCRIPTION NUMBER	presnum	179	7	185	RX_NBR	
22	DATE FILLED	datefill	186	8	193	SERVICED_DTE	YYYYMMDD
23	REIMBURSEMENT DATE	reimbdate	194	8	201	BILL_DTE	YYYYMMDD
24	THERAPEUTIC CLASS CODE QUALIFIER	therapq	202	1	202	'A'	
25	THERAPEUTIC CLASS CODE	therapc	203	17	219	AHFS_CLASS_CDE	
26	THERAPEUTIC CLASS DESCRIPTION	therapd	220	30	249	AHFS_DSC	
27	PLAN REIMBURSEMENT QUALIFIER	planq	250	1	250	'1'	
28	PLAN REIMBURSEMENT AMOUNT	planamt	251	12	262	NET_COST_AMT	9(9)V99-
29	PATIENT LIABILITY AMOUNT	patamt	263	12	274	COPAY_AMT	9(9)V99-
30	NEW/REFILL CODE	refill	275	2	276	RX_REFILL_NBR	9(2)
31	RECORD PURPOSE INDICATOR	recpurp	277	1	277	'M' – MARKET SHARE	
32	REBATE PER UNIT AMOUNT	rebatamt	278	12	289	SPACES	
33	REQUESTED REBATE AMOUNT	rebatreq	290	12	301	SPACES	
34	FORMULARY CODE	formula	302	17	318	FILL_DRUG_FORMULARY_ID	
35	PRESCRIBER ID QUALIFIER	docidq	319	1	319	'D'	
36	PRESCRIBER ID	docid	320	10	329	CLAIM_DEA_NBR	
37	SOURCE CODE	src	330	1	330	RX_COMM_TYPE_CDE : - 'I' - INTERNET - 'M' - MAIL - 'R' - RETAIL	
38	BENEFIT DESIGN	benefit	331	2	332	FORMULARY_TYPE_CDE	
39	CLAIM FORMULARY STATUS	claimstat	333	1	333	FILL_DRUG_FORMULARY_IND	Y/N
40	PLAN FORMULARY LEVEL DATE	plandate	334	3	336	001 - 01/01/2002 Plan	9(3)
41	PLAN FORMULARY LEVEL	planlevel	337	1	337	1 - Managed 2 - Partially Managed 3 - Non-managed	9(1)
42	PRESCRIBER'S STATE	docstate	338	2	339	CLAIM_STATE_CDE	
43	INTERNAL/EXTERNAL INDICATOR	ind	340	1	340	EXTERNAL_SRC_IND - 'I' - INTERNAL - 'E' - EXTERNAL	
44	CLIENT CLASS CODE	class	341	3	343	CLI_CLASS_CDE	9(3)
45	QUARTER INDICATOR	qtr	344	2	345	QTR_IND	
46	CONTRACT ID	contract	346	18	363	CONTRACT_OPERATIONAL_ID	
47	GROUP ID	group	364	18	381	GROUP_OPERATIONAL_ID	
48	PHARMACY CLAIM ID	claimid	382	19	400	PHCY_CLAIM_ID	-9(18)

Figure 2. File layout provided to describe the overall data fields. See Figure 1 for a list of those fields above that are the subject of the assessment.

New York	California
Illinois	Pennsylvania
Michigan	Texas
Massachusetts	Arizona
Florida	

Figure 3. Nine states that are the subject of the assessment. The Dataset draws records from prescriptions filled in these states, collectively termed "States".

2.2. Dataset Geography

The Dataset draws records from prescriptions filled in the states listed in Figure 3, collectively termed the "States". The States reflect differing population models as well as availability of individual-level data and other characteristics and so, results per state may vary.

The Dataset is a *data stream* where the fields and range of associated values remain the same but the values associated with each field are likely to change as prescription information for different patients and pharmacies are included over time.

2.3. HIPAA Privacy Rule

HIPAA dictates allowed disclosures of patient information, specifying who can receive which specific patient data elements and circumstances for sharing patient information with researchers and other organizations not involved in the direct care of the patient.

"HIPAA covered entities" are those organizations that are directly subject to HIPAA. These are organizations directly involved in patient care and/or the billing of patient care. They include: (1) Health plans – HMOs, health insurers, group health plans including employee welfare benefit plans, and agencies administering Medicare, Medicaid, etc.; (2) health care providers; and, (3) clearinghouses and business associates of health plans and providers. Being covered by HIPAA, i.e., being a HIPAA covered entity, carries burdens and responsibilities with violations that can impose criminal and civil liabilities.

With respect to this assessment, managed care organizations and physicians are HIPAA covered entities. Pharmacies are HIPAA covered entities when processing pharmacy claims (but not when paid by the patient direct). Pharmaceutical companies are not HIPAA covered entities. Because the origin of data is billing for patient prescriptions from a HIPAA covered entity, subsequent sharing ("secondary sharing") of the data to the pharmaceutical company must adhere to the provisions set forth in the HIPAA.

The HIPAA Privacy Rule provides three ways in which patient data can be shared free of HIPAA burdens and responsibilities [11]. These are: (1) the Safe Harbor Provision;

(2) the Limited Data Set; and, (3) the Scientific Standard. Collectively, we term these the "HIPAA provisions."

2.3.1 HIPAA Safe Harbor Provision

The HIPAA Safe Harbor provision describes which data elements must be removed in order for the patient-specific data to be shared free of HIPAA burdens and responsibilities. There are 18 categories of patient data elements that must be removed before the data can be shared. These include the following explicitly identifying and demographic fields:

(A) Names;
(B) All geographic subdivisions, except first 3 digits of ZIP code (only 2 digits if ZIP population < 20K)
(C) All elements of dates (except year) for dates
(D) Telephone numbers;
(E) Fax numbers;
(F) Electronic mail addresses;
(G) Social security numbers;
(H) Medical record numbers; and other numbers ...
(N) Web Universal Resource Locators (URLs);
(O) Internet Protocol (IP) address numbers;
(P) Biometric identifiers, etc

If the Dataset would remain useful once these fields are removed, then the HIPAA Safe Harbor Provision would provide the solution. Each managed care organization would remove these fields of data prior to providing the data to pharmaceutical company.

To adhere to the HIPAA Safe Harbor provision, the Dataset would have to make the following changes: (1) prescription_number would have to be removed; and (2) date_of_fill and reimbursement_date could only report the year. All other fields, such as the pharmacy identity would remain the same. We term this version of the Dataset to be "Safe Harbor Data."

2.3.2 HIPAA Limited Data Set

The Limited Data Set provision is for researchers, who enter into a data use agreement and receive the minimal information needed for their research study. An Institutional Review Board decides on the minimum fields and content needed for the specific research protocol provided. Unlike the Safe Harbor provision, full dates and geography may be shared. In this case, the data sharing arrangement between the managed care organizations and the pharmaceutical company is not for research but for marketing, and so, the Limited Data Set provision does not apply.

2.3.3 HIPAA Scientific Standard

The HIPAA Scientific Standard provision allows someone trained in statistics or scientific principles to attest that the provided information has no more than a minimal chance that someone could be re-identified.

Given the nature and content of the health information and based on generally accepted computational, statistical and scientific principles and methods, a person certifies that "the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."

Advantages to using the HIPAA Scientific Standard provision for constructing Dataset are that there is no prohibition against using any particular data elements, no limitation on further sharing, no required data use agreement, and no IRB review. The challenge is to demonstrate that the data elements that comprise Dataset are sufficiently de-identified.

In summary, HIPAA provides three ways to freely share data from HIPAA covered entities –namely, the Safe Harbor provision, the Limited Data Set provision, and the Scientific Standard provision. The Safe Harbor provision requires further redaction of Dataset. The Limited Data Set provision does not apply. Using the Scientific Standard provision on Dataset is the subject of this paper and is further described in the next sections.

3. Methods

3.1 Identifiability

One way to report the risk of re-identification is to determine the number of people to whom a record could refer. This is termed "identifiability." Figure 4 shows two examples in which information is released and compared against a known population. On the left, in Figure 4a, each of the released profiles are ambiguous in terms of head shape and shading. Neither can be uniquely identified. The top released profile matches Hal and Len indistinguishably and the bottom profile ambiguously matches Jim and Mel. The release shown on the right, in Figure 4b is different. There is only one person in the known population (Hal) having the same color and head shape. In this case, the record referring to Hal is uniquely re-identified even though many of Hal's details had been removed.

While unique re-identifications obviously pose a privacy problem, so do situations in which a record maps ambiguously to a few known people. In Figure 4a, both released profiles map to two individuals, but these people are both explicitly known, so they can both be contacted with little effort. Of course, the larger the number of people to whom

a record refers, even if all of the people are known, the greater the effort usually needed to contact so many or make use of the information.

Counting the number of possible re-identifications for a record is a useful measure of privacy risk, but what is needed is a way to estimate the number of people to whom a record might refer.

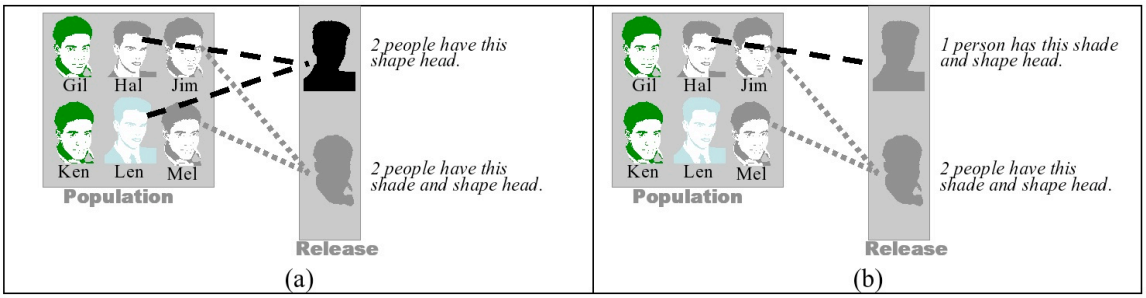

Figure 4. The identifiability of the profiles released in (a) are each ambiguously re-identified to two named persons. The top profile released in (b) is uniquely re-identified to Hal.

3.2 Identifiability of a Dataset

In 2001, Sweeney introduced the Risk Assessment Server as a way to measure the identifiability of a dataset based on her earlier work on identifiability [20]. The Risk Assessment Server has been commercially licensed to companies that continue to use it to report re-identification risks by estimating the number of named persons to which each record could relate given its model of the U.S. population and its knowledge of publicly available datasets [15][16]. The output of the Risk Assessment Server is a plot of identifiability estimates, in graduated size groupings, that report the number of people to which a released record is apt to refer.

Figure 5 shows the results from the Risk Assessment Server based on {date of birth, gender, 5-digit ZIP} in the United States. The lower left plot shows that 87% of the population is uniquely identified by these characteristics. As age information is generalized and as geographical reference to the patient's residence is made less specific, uniqueness deteriorates and privacy protection increases. For example, {year of birth, gender, 5-digit ZIP} drops the unique identifiability to 0.04% (see the lower right plot in Figure 5). This provides some evidence as to why HIPAA Safe Harbor provisions tend to focus on prescribing general demographics, because in general, the unique identifiability of the demographics prescribed by the HIPAA Safe Harbor is 0.04%.

The Results section reports findings from applying the Risk Assessment Server to Dataset. It shows the number of people who could be identified in Dataset, in entirety and by state.

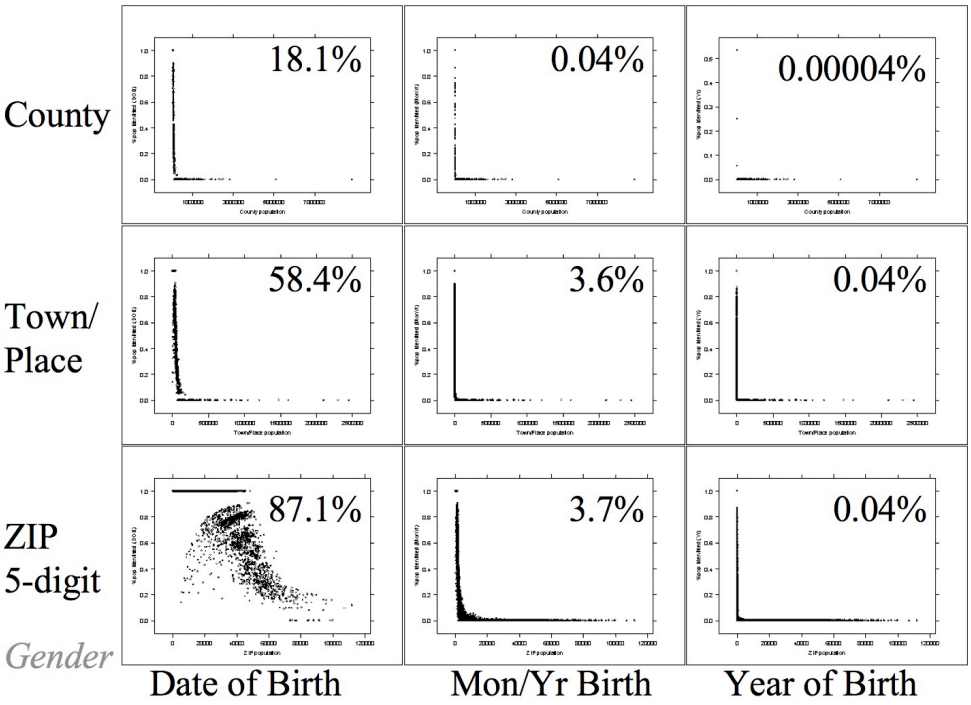

Figure 5. {date of birth, gender, 5-digit ZIP} uniquely identifies 87.1% of USA population, but as ZIP is made less specific, the identifiability drops to 18.1% (bottom to top). Similarly, as the age of the client is made less specific, the identifiability drops to 0.04% (left to right). All values include gender. The horizontal axis of each sub-plot is the number of people who reside in the geographical area and the vertical axis is the percentage of the population uniquely identified by the noted combination of demographics noted. As the demographics are aggregated, the points move towards 0% identifiable. HIPAA Safe Harbor provisions have 0.04% identifiability. See [20].

3.3 Re-identification

A "re-identification" results when a record in the Dataset can likely be related to the patient who is the subject of the record in such a way that direct and reasonably specific communication with the patient (or authorized representative) is possible. This involves determining a likely strategy for learning the explicit name (or SSN) of the patient, or for mailing or phoning the patient's household. Re-identifications include the ability to possibly survey over a small and limited set of named individuals, households, or phone numbers for which the patient is a member of the named set. The Risk Assessment Server not only computes the number of people who could be re-identified but it also reports the re-identification strategy.

4. Identifiability Results

The Risk Assessment Server identifies which fields and/or records in the Dataset are vulnerable to known re-identification inference strategies. The output of the assessment server is a report on the identifiability of the Dataset with respect to those inference strategies. "Identifiability" estimates, in graduated sized groupings, the number of people to which a record is likely to refer. These groupings are called binsizes. For the Dataset, the Risk Assessment Server reports an estimate of how many records match the criteria uniquely (binsize of 1), how many records are likely to relate to one of two possible people (binsize of 2), and so on. In any given report, the number of patients appearing in a smaller binsize is not also counted in a larger binsize, unless otherwise noted as a cumulative result.

In 2003, the pharmaceutical company provided a sample of the Dataset reportedly covered a representative 3-months of transactions in the States ("Sample"). Upon examination of the records in Sample, prescriptions from states beyond the 9 states that are the subject of this engagement were also found. Nevertheless, this engagement remained specifically limited to the 9 states listed above. The Risk Assessment Server evaluates the identifiability of the data stream and not the identifiability of any specific smaller subset of the data stream. The analysis is on the entire Dataset. Below are the results as they appeared in 2003.

Figure 6 reports on the identifiability of records in the Dataset based on the minimal combinations of fields for the State of New York. Binsize 1 accounts for 421,282 (of 17,990,026 or 2.3%) individuals whose information could appear in the Dataset such that their records are likely to provide unique re-identifications. Results for the first 6 binsizes are replicated in an enlarged format in Figure 7. Binsizes 1 and 2 combined account for 1,088,542 (of 17,990,026 or 6.1%) individuals. Binsizes 1 through 6 combined account for 3,298,092 (of 17,990,026 or 18.3%) individuals. Their records are likely to provide unique re-identifications or match up to 6 other identified people.

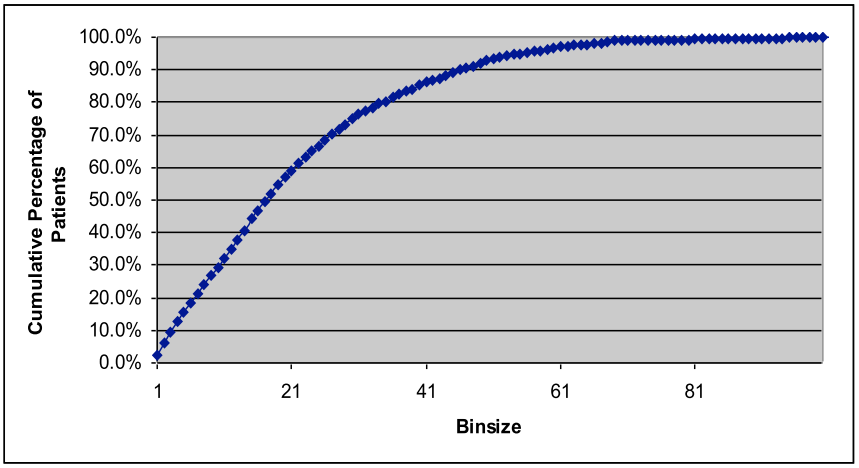

Figure 6. Binsize identifiability rates for the State of New York

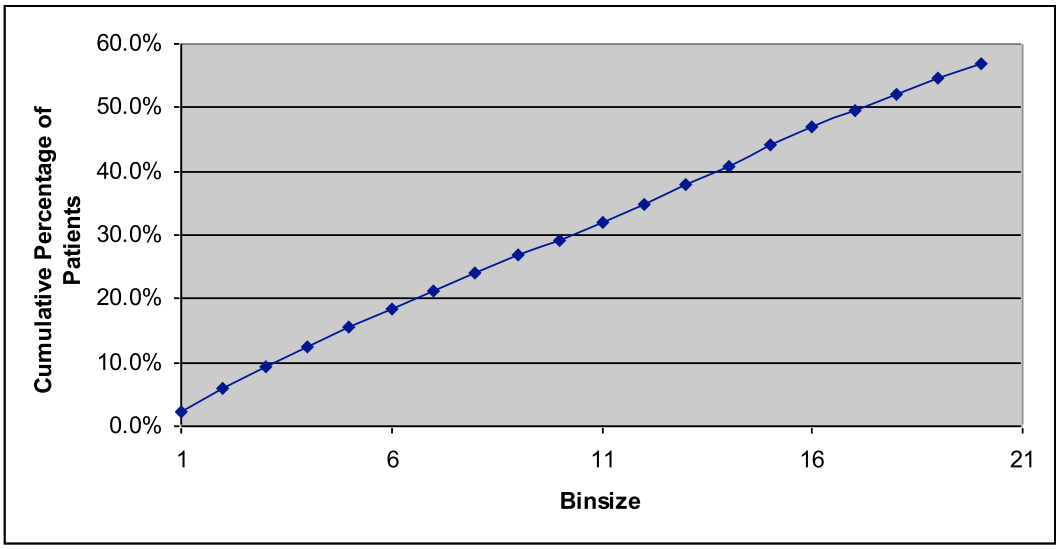

Figure 7. Binsize identifiability rates for the State of New York for bin sizes 1 to 20

Binsize	NY	AZ	TX	PA	IL	CA	FL	MI	MA
1	2.34%	1.24%	0.03%	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%
2	6.05%	3.72%	0.11%	0.10%	0.09%	0.05%	0.01%	0.01%	0.00%
3	9.33%	6.37%	0.24%	0.27%	0.28%	0.10%	0.03%	0.02%	0.01%
4	12.55%	8.93%	0.39%	0.50%	0.61%	0.15%	0.05%	0.05%	0.01%
5	15.48%	11.31%	0.57%	0.79%	1.04%	0.22%	0.09%	0.09%	0.02%
6	18.33%	14.24%	0.77%	1.04%	1.48%	0.29%	0.14%	0.15%	0.03%
7	21.12%	16.93%	0.98%	1.35%	1.94%	0.35%	0.19%	0.23%	0.04%
8	24.11%	19.31%	1.24%	1.74%	2.42%	0.42%	0.24%	0.34%	0.05%
9	26.86%	22.61%	1.50%	2.11%	2.90%	0.50%	0.32%	0.46%	0.07%
10	29.24%	25.98%	1.76%	2.50%	3.38%	0.58%	0.42%	0.61%	0.10%
11	31.94%	28.96%	2.04%	2.99%	3.89%	0.66%	0.51%	0.78%	0.13%
12	34.92%	31.61%	2.36%	3.50%	4.36%	0.75%	0.61%	0.97%	0.16%
13	37.95%	34.49%	2.72%	3.99%	4.77%	0.84%	0.71%	1.16%	0.19%
14	40.62%	37.75%	3.03%	4.54%	5.24%	0.91%	0.83%	1.39%	0.23%
15	44.13%	40.94%	3.40%	5.02%	5.72%	1.00%	0.94%	1.60%	0.27%
16	46.84%	44.29%	3.70%	5.56%	6.20%	1.08%	1.05%	1.77%	0.32%
17	49.43%	48.02%	4.03%	6.19%	6.64%	1.18%	1.18%	1.96%	0.35%
18	52.04%	50.68%	4.38%	6.73%	6.97%	1.28%	1.34%	2.16%	0.39%
19	54.69%	53.33%	4.76%	7.33%	7.27%	1.37%	1.44%	2.38%	0.44%
20	57.01%	55.58%	5.17%	7.90%	7.55%	1.45%	1.62%	2.60%	0.48%

Figure 8 shows the cumulative percentages for each of the states for binsizes 1 through 20.

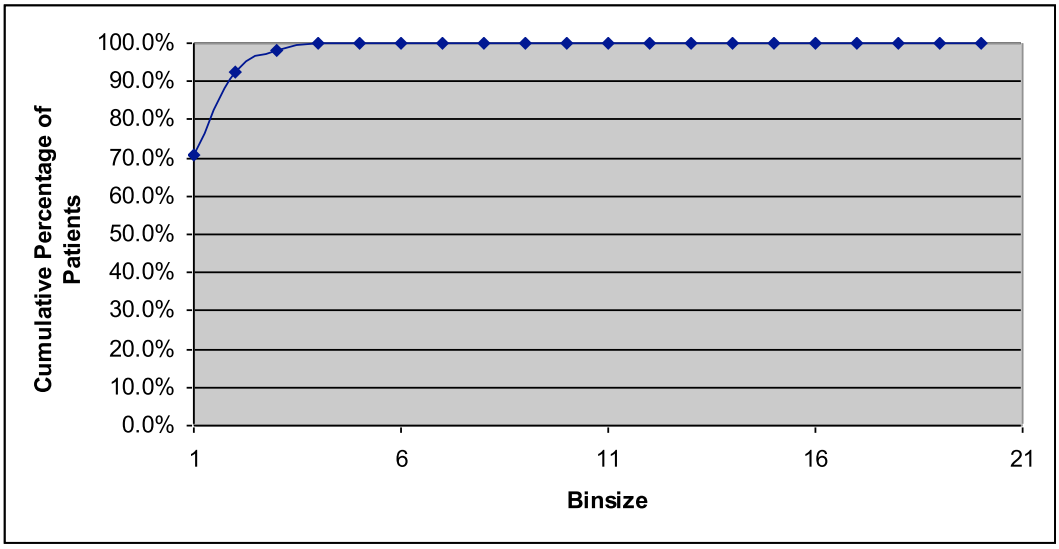

Figure 9. Binsize identifiability rates for the Illinois for bin sizes 1 to 20 for Sample Year

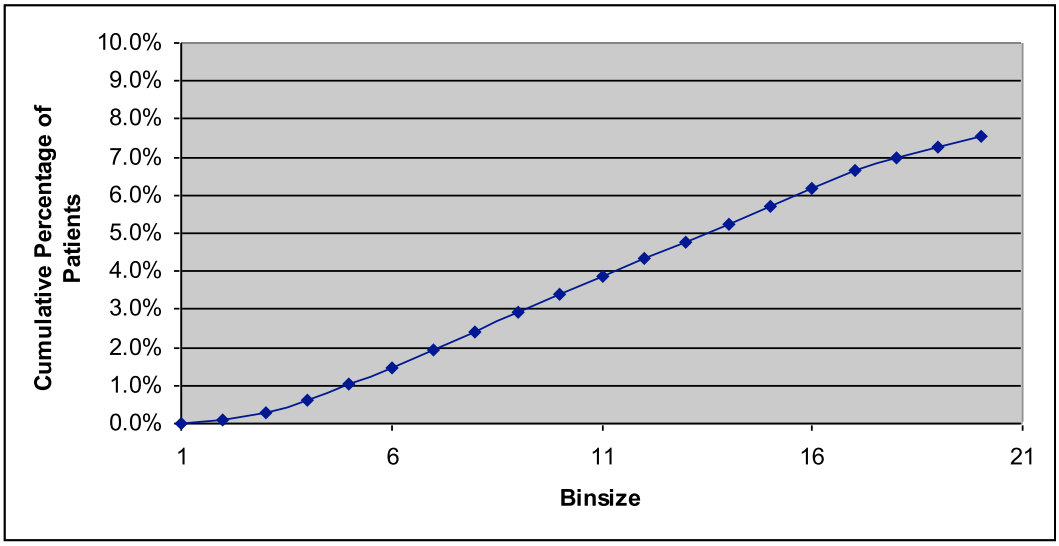

Figure 10. Binsize identifiability rates for the Illinois for bin sizes 1 to 20 (Present)

Figure 9 and Figure 10 show results for Illinois as derived from the Sample provided. The dominant re-identification strategy in Figure 9 uses another prescription dataset that included patient ZIP and age to link to claims data that had the patient's month and year of birth. However, policy changes led to the prescription dataset being removed from public consideration (remaining available for commercial purchase however).

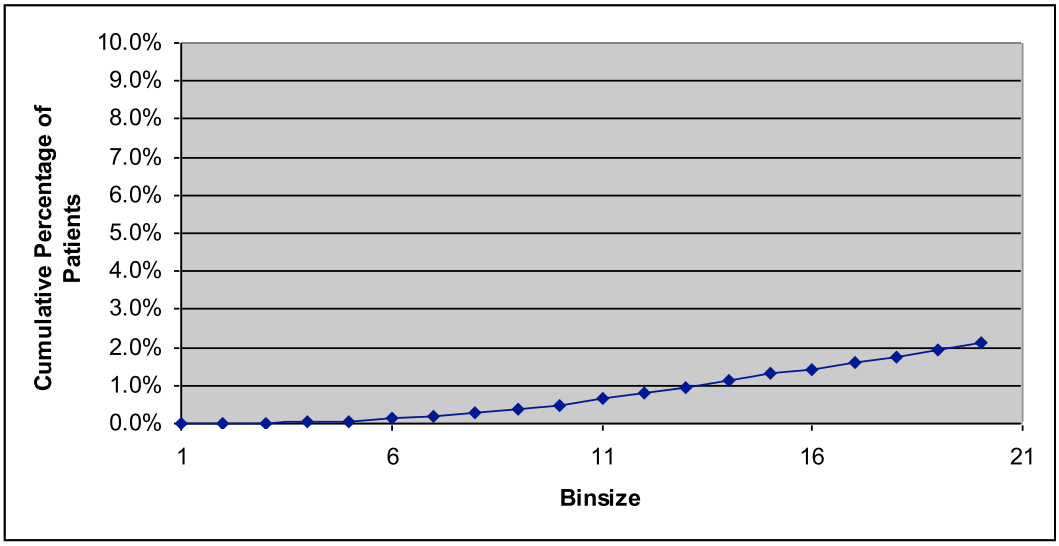

Figure 11. Binsize identifiability rates for the State of Michigan for bin sizes 1 to 20

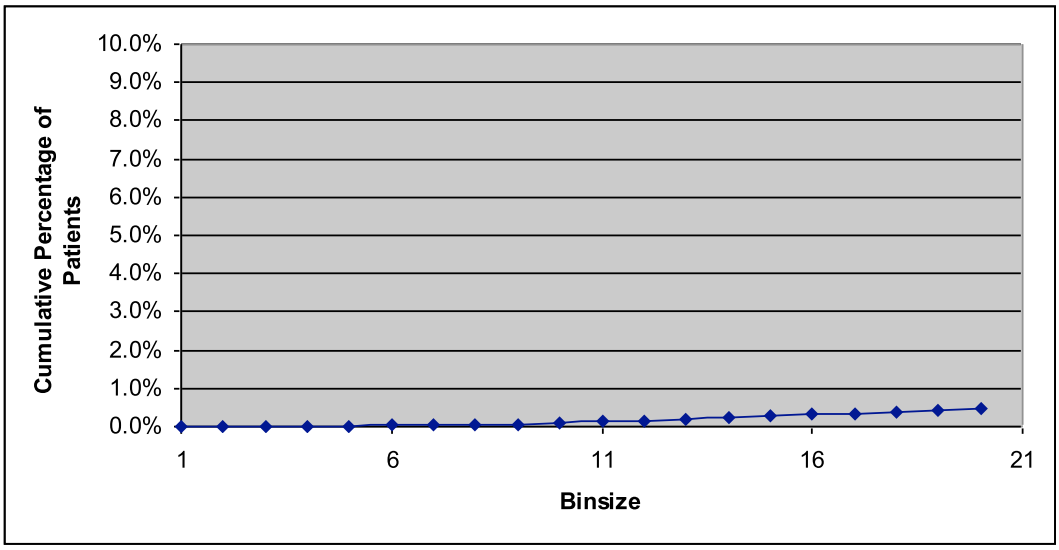

Figure 12. Binsize identifiability rates for the State of Massachusetts for bin sizes 1 to 20

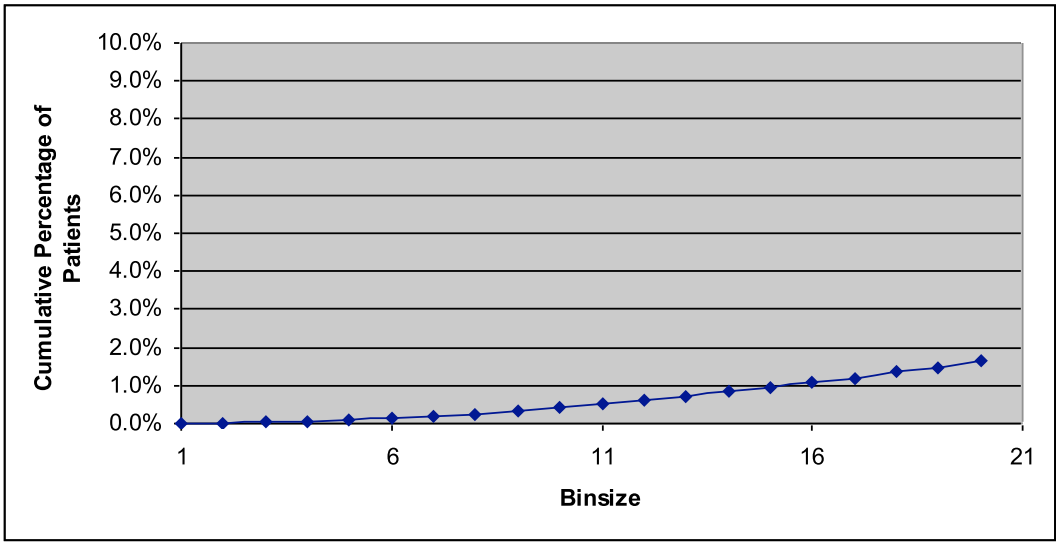

Figure 13. Binsize identifiability rates for the State of Florida for bin sizes 1 to 20

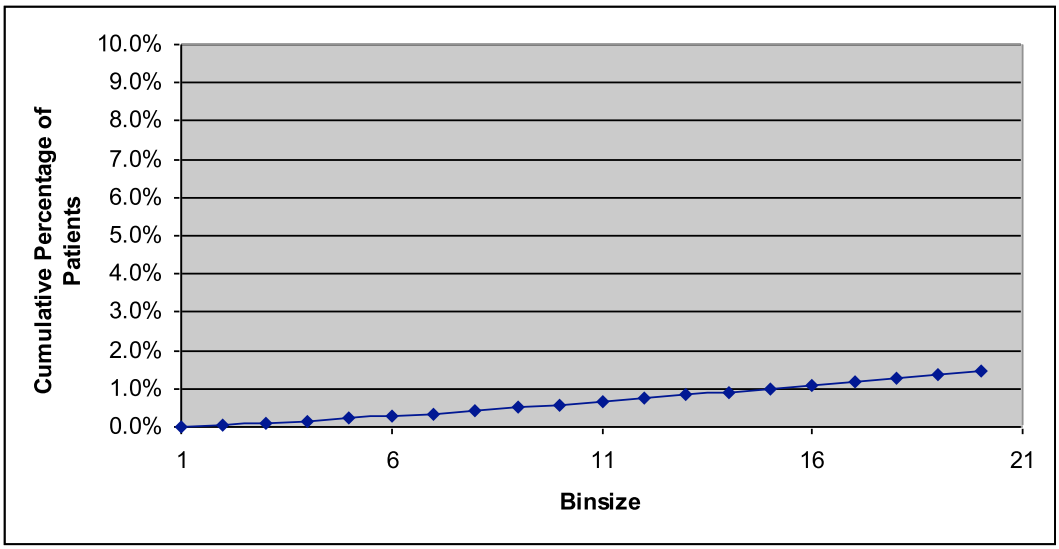

Figure 14. Binsize identifiability rates for the State of California for bin sizes 1 to 20

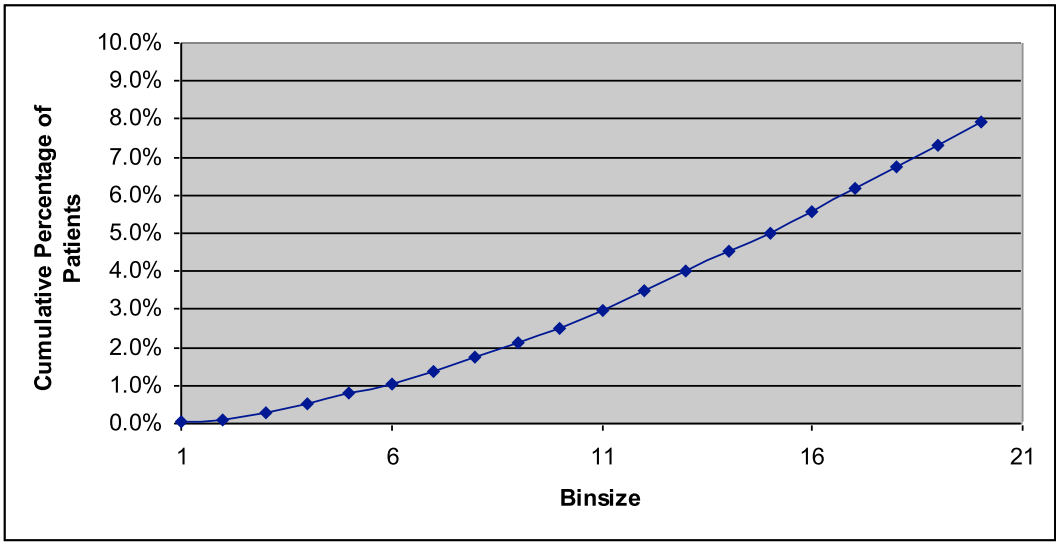

Figure 15. Binsize identifiability rates for the State of Pennsylvania for bin sizes 1 to 20

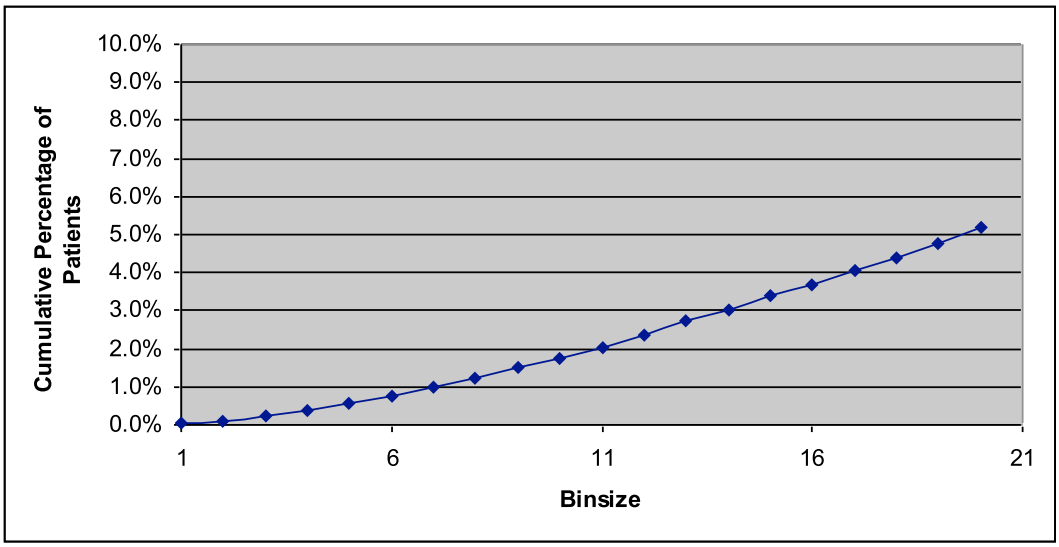

Figure 16. Binsize identifiability rates for the State of Texas for bin sizes 1 to 20

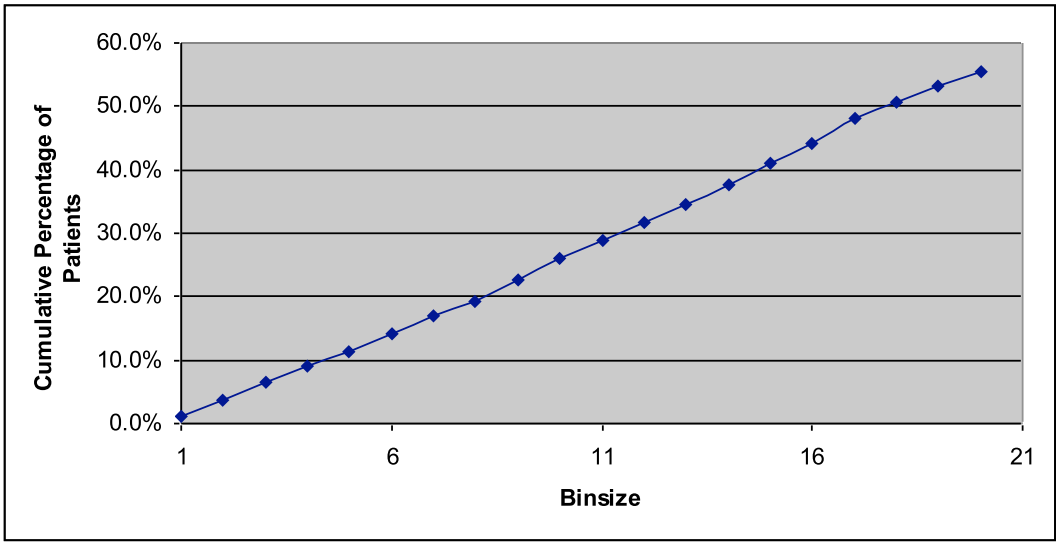

Figure 17. Binsize identifiability rates for the State of Arizona for bin sizes 1 to 20

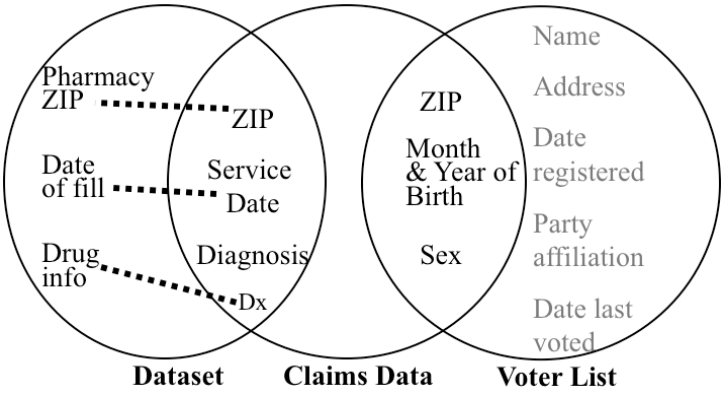

Figure 18. Depiction of common re-identification strategy that provided re-identification results.

The principal re-identification strategy identified by the Risk Assessment Server is depicted in Figure 18. It is a 2-stage re-identification that uses publicly available hospital discharge and ambulatory claims data (e.g. described in [21]) to learn more patient demographics. The patient demographics are then linked to a population register, such as a voter list to identify the patients by name.

In order to facilitate the linking, we used additional prescription data to construct models for inferring patient ZIP from pharmacy ZIP, service date from the date of fill and prescription information, and we used databases that relate medications to diseases (e.g., [22]).

Because the availability of claims data and the fields contained within claims data varies from state to state, as well as differences in the population demographics in each state, the results vary from state to state.

Below are some additional observations.

Observation #1. Personnel Information.
Among the benefits plans explicitly listed under the plan name field in the Dataset is the pharmaceutical company itself. The prescriptions of the company's personnel appear in the Dataset. Using information held semi-privately or privately within the company, the identities of these people could be reliably determined and therefore as an employer, the company could have access to sensitive health information on its employees. This is not the expressed use of the Dataset, but is an important matter. In the reported results above and discussion that follows, the company's personnel re-identifications are not included.

Observation #2. Targeted groups are sensitive.
Reliable re-identifications of targeted groups of people are possible in the Dataset. While the total quantities of people re-identified by a single targeted attack may not itself provide large numbers of re-identifications, the fact that most or many people in the targeted group can be re-identified is of concern. A glaring example in the Sample was patients with HIV.

5. Discussion

Knowing that people can be re-identified in the data means it is not anonymous, but it is sufficiently de-identified? For years, the HIPAA Scientific Standard was crippled by this, but eventually a new scientific-legal approached emerged based on counting the number of people who could be re-identified in the data. Below is a description of this method and its comparison to other standards.

5.1 Compliance with the HIPAA Scientific Standard

Under the HIPAA Scientific Standard, the risk for re-identification has to be "very small" but the regulation never provides any explicit means to quantify how small is very small. So, lawyers and statisticians alike were leery to use the provision. Sweeney introduced the Privacert Risk Assessment model for HIPAA Compliance ("Privacert Model") as a way of determining whether data are sufficiently de-identified under the HIPAA Scientific Standard [16]. The idea is simple: accept a dataset that doesn't make any more people identifiable than is made identifiable by the HIPAA Safe Harbor. Recall earlier, that in general the identifiability of the HIPAA Safe Harbor is 0.04%. The Privacert Model therefore, in general, accepts a dataset that may include fields not allowed by the HIPAA Safe Harbor (e.g., full dates and ZIP codes) provided no more people are put at risk to re-identification than would be allowed by the HIPAA Safe Harbor.

Qunitles was the first to use a version of the approach in real-world practice [17] and bioterrorism surveillance efforts sought to use the approach more widely. Over the last 6 years, the approach has been used commercially by numerous large insurance and data mining companies and government agencies [16].

Applying the Privacert Model generically to the results in Figure 8, data from the states of New York and Arizona would not satisfy the standard where as data in the states of Texas, Pennsylvania, Illinois, California, Florida, Michigan, and Massachusetts would.

In 2003, when these identifiability results were first produced, Stoto conducted a survey of statistical literature and inferred that in comparison to the practices of federal statistical offices an acceptable standard might be 4% [14]. But these offices are operating with different kinds of data under different regulatory regimes. In comparison, the 0.04% standard described earlier is derived from risks prescribed by HIPAA itself.

Observation: These results are based on the identifiability of isolated claims. If the claims are combined for the same person over time, the combination would likely provide significantly more unique re-identifications as the prescription history over time correlates with new diagnoses over time.

5.2 De-identification Failure

In closing, this writing simultaneously shows the best of de-identification practice and by doing so, exposes the worst of de-identification as a regime. The best is that HIPAA provides a mechanism in which privacy risk from de-identification can be assessed relative to its own standard (the Safe Harbor provision). Notwithstanding this solution, de-identification is fraught with serious perils that demand a new approach. Below are four concerns.

Dynamic risk.
Data sufficiently de-identified today may be re-identifiable tomorrow because there is no knowledge or coordination of datasets that may be available tomorrow. For example, access to data linkable to prescription claims in Illinois changed radically during the study period. See Figure 9 and Figure 10. In this case, the re-identification dropped from what would have been the most identifiable state (71% unique identification) to one of the least identifiable states (0.01% unique identification) due to the redaction of fields and changes in data access policies to prescription and other data. As more data is made readily available, such as credit card purchases, online prescription purchases, email messages about refills, and cell phone location data, re-identification risks increase because there is no coordination between the data releases. There is not necessarily coordination between other releases of prescription data made by the same managed care organizations.

Undisclosed risk.
A data recipient could hold other data which could facilitate linking but not have been considered in the risk assessment or that could have been acquired after the risk

assessment completed. The results in Figure 9 and Figure 10 provide an example. The pharmaceutical company that is the subject of the assessment of Dataset could have access to the other prescription dataset referred that holds more identifiable patient data (see Figure 9), and could then use that dataset to re-identify patients in Dataset. A company falsely believing the data could not be re-identified may unknowingly put data at risk or not seek necessary security precautions. After all, data that adheres to the HIPAA Safe Harbor or Scientific Standard provisions can be shared freely without further review or restrictions on use. Over time, companies relying on data streams of de-identified data can easily fall victim to supply chain problems due to unforeseen fluctuations in re-identification risk.

Lack of universal specification of fields.
The application of a level of acceptable uniqueness (e.g., 0.04%) is not aligned with a universal list of data elements. Re-identification risks vary with population demographics and data availability, requiring data elements to vary across populations in order to achieve the same level of risk. Holding the level of uniqueness constant, requires changes in fields across states. For example, to get Dataset to be compliant with the 0.04% standard would require further redaction in the fields for New York (2.34%), but no further redaction in the fields for Massachusetts (0.0%).

Lack of accountability.
A person could be egregiously harmed by data sharing, but not be able to show the hidden trail that led to the harm. There is no way for a person to know who holds data that could be re-identified to him at any given time. Even if he knows direct data recipients, such as the pharmacy he visits, he cannot necessarily know about other recipients, such as the pharmaceutical company sending him direct ads. This lack of transparency provides a lack of accountability in tracking harms. An example is the compilation and use of personal prescription profiles by companies [23].

What is needed are strong technical solutions with guarantees that no on can be re-identified yet the information remain practically useful. Such approaches offer anonymity by demonstrating that records cannot even be reliably matched back to the source from which they come. Imagine Dataset having a computer program modify its values so that each record ambiguously relates to the original data. If the records cannot reliably match to the original source, then the protection holds regardless of any other external sources of information. Examples of such technical approaches to anonymization are k-anonymity [9] and differential privacy [10]. Unfortunately, there has been little or no incentive to consider these approaches because of the availability of de-identified (but often re-identifiable) data.

Another technical approach (termed “multiparty computation”) leverages the use of network communication so that data holders do not share person-specific data but instead jointly compute desired results (see [24] for a real-world example using homeless service utilization). To apply this approach to Dataset, each managed care organization would have a computer that jointly aggregated prescription information by disease over time and by pharmacy location over time and then provide the aggregated

results to the pharmaceutical company. The company would not receive patient-level details. These kinds of privacy solutions (data anonymization and multiparty computations) use technology to provide necessary protection for businesses and individuals, but as long as inferior de-identification standards exist, they will not flourish and Americans and American companies will not have privacy and utility, but will have to wrongfully choose between privacy or utility.

References

1 PriceWaterhouseCoopers. Transforming healthcare through secondary use of health data. 2009.

2 IMS Health. IMS Facts at a Glance. As of September 30, 2010, http://www.imshealth.com/

3 Acxiom. FAQs and EEOC Guidelines. As of Septemer 30, 2010 http://www.acxiom.com/products_and_services/background_screening/faq/Pages/FAQs.aspx

4 MedMining. Welcome to MedMining. As of September 30, 2010 http://www.medmining.com/

5 Barbarao M and Zeller T. A face is exposed for AOL searcher 4417749. New York Times. August 9, 2006, Page A1.

6 Narayanan A and Shmatikov V. Robust de-anonymization of large sparse datasets [Netflix]. IEEE Symposium on Research in Security and Privacy, Oakland, CA, 2008.

7 Felch J. DNA databases blocked from the public. Los Angeles Times. August 29 2008, PageA31.

8 Sweeney L. Weaving technology and policy together to maintain confidentiality. Journal of Law. Medicine and Ethics, volume 25, 1997 http://dataprivacylab.org/dataprivacy/projects/law/law1.html

9 Sweeney L. k-anonymity: a model for protecting privacy. In International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, volume 10, 2002. http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.html

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

11 45 CFR 164 Health Insurance Portability and Accountability Act (HIPAA) of 1996 (P.L. 104-191)

12 Latanya Sweeney. http://dataprivacylab.org/people/sweeney/index.html

13 Michael Stoto. http://explore.georgetown.edu/people/stotom/?PageTemplateID=179

14 Stoto, M. The Identifiability of Pharmaceutical Data: a Test of the Statistical Alternative to HIPAA's Safe Harbor. In CD-only annex to Domingo-Ferrer J, Franconi L, eds. Privacy in Statistical Databases, Lecture Notes in Computer Science 4302, Springer, 2006.

15 Sweeney, L. Data Sharing Under HIPAA: 12 Years Later. Invited presentation to the HHS Workshop on the HIPAA Privacy Rule's De-Identification Standard, Office of Civil Rights, U.S. Dept. of Health and Human Services, Washington, DC. March 8, 2010. http://hhshipaaprivacy.com/assets/5/resources/Panel2_Sweeney.pdf

16 Privacert Risk Assessment Server (licensed to Privacert, Inc. by L. Sweeney, Carnegie Mellon University). http://privacert.com/assess/index.html

17 Beach, J.Health Care Databases under HIPAA: Statistical Approaches to De-identification of Protected Health Information. DIMACS presentation. December 10, 2003. http://dimacs.rutgers.edu/Workshops/Health/abstracts.html and http://www.zurich.ibm.com/pdf/privacy/report3-final.pdf

18 K. Benitez, G. Loukides, and B. Malin. Beyond Safe Harbor: automatic discovery of health information de-identification policy alternatives. In Proceedings of the ACM International Health Informatics Symposium (IHI). 2010: 163-172.

19 B. Malin, K. Benitez, and D. Masys. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. Journal of the American Medical Informatics Association. 2011; 18(1): 3-10.

20 Sweeney, L. Uniqueness of Simple Demographics in the U.S. Population. Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report LIDAP-WP4. Pittsburgh: 2000. Shorter version available as: Simple Demographics Often Identify People Uniquely. Working Paper 2. 2000. http://dataprivacylab.org/projects/identifiability/index.html

21 Essig, C. Illinois rakes in millions selling personal data. Pantagram.com. April 16, 2010. http://www.pantagraph.com/news/state-and-regional/illinois/article_370b913e-4991-11df-ac1b-001cc4c002e0.html

22 Rx List. See online exemplar version at http://www.rxlist.com/script/main/hp.asp

23 "And You Thought a Prescription Was Private". The New York Times, August 9, 2009. See also https://www.annualmedicalreport.com/personal-prescription-data-is-bought-and-sold-by-health-insurers-pharmaceutical-companies/

24 Sweeney, L. Demonstration of a privacy-preserving system that performs an unduplicated accounting of services across homeless programs. Data Privacy Lab Working Paper 902. October 2007. http://dataprivacylab.org/projects/homeless/index2.html