HHS Public Access

Author manuscript

Technol Sci. Author manuscript; available in PMC 2019 January 23.

Published in final edited form as:

Technol Sci. 2017 ; 2017: .

Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study

Latanya Sweeney[1], **Ji Su Yoo**[1], **Laura Perovich**[4], **Katherine E. Boronow**[2], **Phil Brown**[3], and **Julia Green Brody**[2]

[1]Harvard University, Cambridge, MA

[2]Silent Spring Institute, Newton, MA

[3]Northeastern University, Boston, MA

[4]MIT Media Lab, Cambridge, MA

Abstract

Researchers are increasingly asked to share research data as part of publication and funding processes and to maximize the benefits of publicly funded research. The Safe Harbor provision of the U.S. Health Information Portability and Accountability Act (HIPAA) offers guidance to researchers by prescribing how to redact data for public sharing. For example, the provision requires removing explicit identifiers (such as name, address and other personally identifiable information), reporting dates in years, and reducing some or all digits of a postal (or ZIP) code. Is this sufficient? Can research participants still be re-identified in research data that adhere to the HIPAA Safe Harbor standard? In 2006, researchers collected air and dust samples and interviewed residents of 50 homes from Bolinas and Richmond (Atchison Village and Liberty Village), California, to analyze the residents' exposure to pollutants. The study, known as the Northern California Household Exposure Study [1], led to publications that have been cited hundreds of times. We conducted experiments with separate "attacker" and "scorer" teams to see whether we could identify study participants from two versions of the data redacted beyond the HIPAA standard, one in which all dates were reported in ranges of 10 or 20 years and another in which a study participant's birth year was reported exactly. The attackers were blinded to the names and addresses of the participants, and the scorers were blinded to the strategy.

Results summary:

We correctly distinguished the 10 records from Bolinas and 32 records from Atchison Village, and we presented 9 records that included the 8 correct records from Liberty Village. When the redacted data contained the exact birth year, as allowed by HIPAA Safe Harbor, we correctly identified 8 of 32 (25 percent) Atchison Village participants by name and 9 of 32 (28 percent) by address. In comparison, earlier studies found unique re-identification rates in data that adhered to the level prescribed by HIPAA Safe Harbor to be much lower, namely 0.013 percent [2] and 0.04 percent [3]. However, these earlier studies relied solely on demographic fields for re-identification. Our experiments used fields beyond demographics (e.g., housing characteristics), and by doing so, substantially increased re-identification risk in data compliant with HIPAA Safe Harbor. Even in more heavily redacted data showing participants' birth years in 10- or 20-year ranges, we uniquely and correctly identified 1 of 32 (3 percent) of the Atchison Village study participants by name and

address and identified 4 of 32 (13 percent) participants as being one of fewer than five named choices. No correct results were found for Liberty Village or Bolinas under these conditions. These results suggest that the HIPAA Safe Harbor is not a sufficient privacy guard for environmental health data and bring into question the practice of using the HIPAA Safe Harbor standard as a general rule for “de-identifying” other datasets in today’s data-rich, networked environment.

Abstract

Re-identification strategy to associate an ID in the Study Data with an Address and Name of a participant in the study

Introduction

The Privacy Rule of the Health Information Portability and Accountability Act (HIPAA) is the U.S. federal regulation that governs the sharing of patient medical information by doctors, hospitals, and others involved in direct patient care or in the billing for that care [4]. Improper handling of patient information can result in civil and criminal penalties. For example, an incidental data breach could cost $50,000, and patient information knowingly disclosed could result in a criminal penalty of $250,000 and ten years of imprisonment [5].

On the other hand, if the data are redacted as prescribed by the Safe Harbor provision within HIPAA, then the redacted version can be shared freely without concern for civil or criminal penalties [6]. HIPAA Safe Harbor requires eliminating 16 kinds of patient identifiers (e.g., patient name, Social Security number, email address, and telephone, account, and all other record numbers) and generalizing date and geography information: dates must be reported as years, and the smallest reportable geographic subdivision is the first 3 digits of the ZIP (postal) code (unless the three-digit ZIP code contains fewer than 20,000 people, in which case it is reported as 000) [7]. Personal health information redacted in this format can be shared widely, online or offline, with no restrictions and without a data use agreement. Promulgated on August 14, 2002, the HIPAA Privacy Rule remains in effect today. Although it formally applies to patient health records, HIPAA Safe Harbor is sometimes proposed as a benchmark in other contexts, such as Institutional Review Board oversight of research [8].

The HIPAA Safe Harbor standard uses a traditional pillar of data privacy known as de-identification – the removal of explicit identifiers from data to make the result sufficiently anonymous. The rationale behind de-identification is simple. If an individual cannot be distinctly identified in data, then no individual’s privacy interests are affected, so the data can be shared widely for many worthy purposes.

HIPAA Safe Harbor is convenient. A researcher can easily comply by merely making the appropriate data redactions. No special computer programs, statistical modeling, or advanced analysis is necessary. But does the HIPAA Safe Harbor adequately protect privacy?

Re-identification

When sharing personal data widely, the biggest privacy threat to “de-identified” data is “re-identification” – the ability for an interested adversary to use reasonable effort to match details in the de-identified dataset to distinct persons sufficiently to contact them. We use the term “named person” to refer to having sufficient information to identify a person by name and “named location” to refer to having sufficient information to identify a physical place having few people. An example of a named location is the residential address of a family. If specific records in a de-identified dataset can be associated with one or few named people or named locations, then we say in this writing that the dataset is re-identified (regardless of whether the associated records contain the true identity). Harm from a re-identification may result if sensitive information contained in the data becomes known about named persons or named locations. For example, when Sweeney re-identified hospital discharge data released by Washington state, her re-identification exposed records that included sensitive information such as “references to venereal diseases, drug dependency, alcohol use, [and] tobacco use” [9].

A “unique re-identification” occurs when a record in the data matches exactly one named person or location. A “group re-identification” occurs when one or a few records in the dataset match a small number of named people or locations. Both unique and group re-identifications raise privacy concerns. For example, if a de-identified dataset does not include names or home addresses, but does include age in months, gender, and 5-digit ZIP codes, it is possible to use publicly available websites to deduce the identity and home addresses of many individuals in the database, as was demonstrated by the recent re-identification of de-identified medical records from Washington state [9]. A one-to-few match can be just as damaging as a one-to-one match. For example, showing that a record in a de-identified dataset of lead poisoning cases belongs to one of few named locations could cause all the real estate properties in the group to suffer adverse consequences, even though only one of the named locations actually has the lead poisoning risk. As another example, a group re-identification of de-identified medical records showing that 6 of 7 named people have a genetic disposition toward cancer would leave the impression that each individual was equally likely (6 in 7) to have that condition, including the individual without the condition. It is well recognized that one-to-few and few-to-few re-identifications pose privacy risks similar to unique re-identification [10].

Rarely is zero risk of re-identification required in publicly shared datasets, and HIPAA is no exception. In 2011 El Emam et al. conducted a review of 14 published re-identification attacks [11] Of the 14 examples, the authors dismiss 11 as being conducted by researchers solely to demonstrate or evaluate the existence of a risk of re-identification, not necessarily knowing whether the re-identification was correct. They classify the work of Narayanan and Shmatikov [12] in this category. Narayanan and Shmatikov demonstrated the possibility of re-identifying published Netflix rental histories from the (identified) movie reviews submitted by Netflix customers.

More generally, Sweeney used 1990 Census data to estimate that 0.04 percent of the United States population was uniquely identified by the basic demographic fields allowed by the

HIPAA Safe Harbor – namely, year of birth, gender, and first 3 digits of ZIP [3]. Both the study by Kwok and Lafky and the study by Sweeney examined only demographic fields, and both found low likelihoods for unique re-identifications. Are we failing to consider other possible risks of re-identification by only studying those addressed by HIPAA Safe Harbor? What about small group re-identifications? What about matching on fields other than demographics?

Answers to these questions are critical as researchers seek to share research data widely. Many academic publications now require authors to submit a version of the data on which results are reported as a condition of publication (for examples, see [13, 14]). Also, federal, state and city governments increasingly make their datasets publicly available as part of open data initiatives (e.g., [15]). Sharing research data freely is important for science because it allows other researchers to verify published findings and can lower research costs through data reuse. U.S. regulations for data sharing are sector–specific, and most kinds of research data are not subject to any federal data sharing standards. In cases where the data are not subject to HIPAA, researchers and Institutional Review Boards that approve data sharing for research often wonder whether the HIPAA Safe Harbor's prescriptive solution will suffice [8]. In what cases does the HIPAA Safe Harbor provide sufficient privacy protections for sharing research data?

Background

In 2006, researchers (including this paper's authors Brody, Perovich, and Brown) from the Silent Spring Institute, Northeastern University, and the University of California, Berkeley, with funding from the National Institute of Environmental Health Sciences, collected air and dust samples and interviewed residents in 40 homes in two neighborhoods in Richmond, California, and 10 homes in Bolinas, California. This project, known as the Northern California Household Exposure Study (HES), aimed to improve scientific understanding of indoor exposures to pollutants [1]. The two communities of Atchison and Liberty Villages in Richmond were chosen for the study because they were industrial communities within a few miles of the Chevron Richmond Refinery, major transportation corridors, and a marine port [14] Bolinas was chosen on the advice of the community advisory council to provide a rural comparison within the same region. The researchers published findings in leading journals with summary statistics that describe the demographics of the research participants and detailed analysis of chemical pollutants found in the participants' homes and outdoor air [1, 16, 17, 18, 19].

In addition, the researchers wanted to share the study data widely for further analysis by others. For example, the US Environmental Protection Agency requested access to the HES data to estimate human exposures from consumer products. However, the researchers sought to honor the privacy statement made to research participants when sharing data. The informed consent for the study states:

HOW WILL THE DATA BE KEPT CONFIDENTIAL?

All information that could identify you will remain confidential to the full extent of the law. The samples from your home will be identified with a number rather than your name when

they are sent to the laboratory for tests. Any record that includes your name or personal identifying information will be kept in locked file cabinets and access to these records will be restricted to researchers involved in this study.

In research studies like these, the risk of re-identification is a matter not only of privacy protections required by the IRB and described to participants when they first consented to the study, but also of researchers' broader responsibilities to avoid harms to participants. For example, if participants' names or addresses can be matched to the research data, the information about certain pollutants in their homes could adversely impact the value of their properties. If participants were renters, identification might lead landlords to terminate or refuse to renew leases in the belief that the renters may have exposed them to economic problems by participating in research.

Knowledge from an attempt to re-identify a HIPAA Safe Harbor–compliant version of the HES research data can inform data-sharing practices, informed-consent documents, and the development of new strategies to protect study participants.

HES researchers did not collect the data as part of patient medical care, so it was not subject to HIPAA. Still, they faced decisions about data redaction and wondered what protection a HIPAA Safe Harbor version would offer.

In the next sections, we report on our attempt to match names and addresses of research participants to a HIPAA Safe Harbor–compliant version of the demographic and house information collected as part of the exposure study. The HES researchers have never shared the data publicly, so this experiment reports on risks of data that would result if the data were shared in compliance with HIPAA Safe Harbor.

About Re-identification

A "re-identification strategy" in this writing is a means to assign identifying information to entities (e.g., people or addresses) whose information is believed to appear in de-identified records. Approaches typically include a stepwise process applied to various datasets, where one of the datasets is the de-identified dataset itself.

The relevant outcome of a re-identification strategy is usually a set of sufficiently small group re-identifications. The total "number of re-identifications" is the number of records re-identified, regardless of whether the correct identification is included. If only unique re-identifications are of interest, then the number of re-identifications is the number of one-to-one associations found. When larger-sized groups are relevant, then the number of re-identifications of records in the dataset is the number of groups. For example, consider a re-identification having 4 groups, with 2 named people in each group. One person in each of the two person groups is believed to be the correct person, but the re-identification strategy does not distinguish which of the two named people that person might be. Therefore, the number of re-identifications is 4, one person from each group.

A re-identification does not necessarily need to be correct to be harmful. If a sufficiently reliable re-identification strategy strongly associates a record to a person, then that person

will likely suffer the same harm whether they are named correctly or incorrectly. We use the term “correct re-identification” to distinguish instances when re-identification identifies the true person. Consistent with data privacy literature, both re-identifications and correct re-identifications are important.

In prior work, Sweeney introduced the notion of a “binsize” as the number of entities (people or addresses) that matches one or more de-identified records indistinguishably [9, 20, 21]. Unique re-identifications have a binsize of 1, denoting a single one-to-one matchup, uniquely identifying the person or address. A binsize of k lists k possible matches to a single person or address.

The number of unique re-identifications is the value at binsize 1 (we write k=1). Past government data-sharing policies required suppression of data that could lead to re-identifications for binsizes less than 5 ($k<5$) (e.g., [22]). Recent government data-sharing policies proscribes re-identifications for binsizes less than 11 ($k<11$) (e.g., [23]). Guidelines for defamation cases have focused on expecting no re-identifications for binsizes less than 20 ($k<20$) (e.g., [24]) or 25 (e.g., [25, 26]). Therefore, for generalizability, we report the number of re-identifications having thresholds at $k=1$, $k<5$, $k<11$, and $k<20$.

A re-identification strategy identifies a “risk pool” for groups 1 to k [27], comprising all distinct entities named in the re-identified groups from size 1 to k. Risk pools are important because they identify which other entities may be harmed indiscriminately. In the prior example in which the results of a re-identification strategy was 4 groups with two named people for each group, then 8 named people are in the risk pool and the total number of re-identifications is 4. Notice that the risk pool, as defined here, relates to a re-identification strategy. Another re-identification strategy operating on the same de-identified dataset may generate a different risk pool.

Methods

We split ourselves into two separate teams, the “Scorers” and the “Attackers,” to conduct an experiment in which the names and addresses of study participants, held by the Scorers, were kept private from the Attackers, and re-identification strategies, developed and conducted by the Attackers, were kept private from the Scorers until the experiment’s conclusion. Although we met to organize ourselves, actual names, addresses and re-identification strategies were not shared during these discussions. The Attackers attempted to put names and addresses to records in a HIPAA Safe Harbor–compliant dataset and then submitted batched matches to the Scorers. The Scorers consisted of co-authors Brody, Perovich, Boronow, and Brown. The Attackers consisted of co-authors Sweeney and Yoo. The Attackers performed two preliminary iterations with the Scorers before establishing the more succinct version of the re-identification strategy described here.

A re-identification experiment requires registers containing named people and locations to match to the de-identified records. Because the HES is a study of air and dust samples from homes, we used property tax registers for Atchison Village and Bolinas, California, where most homes were owned by residents. Liberty Village is rental housing, so an address

register had to be constructed. HES participants lived in the homes tested, so registers were constructed of the names and addresses of adult residents in those communities during the study period. Below is a description of materials, subjects, and our 7-step approach.

Materials

The “HES Study Dataset” refers to the original data collected in the Northern California Household Exposure Study (HES) [1]. These data include demographic information about participants such as race, gender, birthdate, education level, the year they moved into their residence, and whether they owned the home. Information about residences includes room descriptions and dimensions, use of carpet per room, year the house was built, heating and cooking options, and numerous details about appliances, cleaning choices, pesticide use, pets, and lawn care. The HES Study Dataset also includes extensive one-time air and dust measurements taken at each home. The original dataset was redacted and modified to comply with HIPAA prior to re-identification (see Detailed Approach).

“HES Publications” refers to the set of papers previously published about the study (including [1, 16, 17, 18, 19, 28, 29]).

The “Atchison Village Property Register” is a copy of the 2006 tax assessor data for Atchison Village purchased from the County of Contra Costa Assessor’s Office for $35 [29]. For each homeowner in Atchison Village, these data include the names of the property owners, the address of the property, the number of rooms, baths and bedrooms, the date the house was built, whether a garage is present, and the total land area.

The “Bolinas Tax Data” is a copy of the 2006 tax assessor data for Bolinas, California purchased for $112 from the Marin County Assessor’s office [40]. Unlike the Atchison Village Property Register, the Bolinas tax data do not include any of the specific housing characteristics, such as the number of rooms, baths and bedrooms, the date the house was built, whether a garage is present, and the total land area. The data do include the names of the owners, the address of the property, and various tax parameters.

An online subscription to a “data broker” website allows searches that associate the names and biographical information of people to known addresses, and vice versa, and therefore provided the ability for the Attackers to construct a register of people who live in a particular geographical area. Hundreds of data brokers sell personal information on Americans, including names, telephone numbers, birthdates, and current and historical addresses [31, 32, 33]. Some data brokers list the dates at which the person was known to reside at an address. Many data broker websites allow searches by any field – e.g., by name or historic or current address. Subscription costs typically range from $12.95 per month to $99 per month for an unlimited number of searches or $1 to $5 per search.

Internet tools include a web browser and the use of Google Earth and Google Street View images [34]. Additional data include the Census data on the popularity of occurrences of first names by gender (“Gender Names”) [35] and of last names by race and ethnicity, specifically Black, White, Asian, Native American, and Hispanic surnames (“Race Names”)

[36]. Computational tools include a spreadsheet program, a text editor program, and the Python programming language [37] running on off-the-shelf laptops.

Subjects

The subjects of the re-identification experiment were the 50 adult participants in the HES study, 10 of whom resided in Bolinas, California, and 40 of whom resided in the Atchison Village or Liberty Village communities of Richmond, California.

Approach

The Attackers crafted a re-identification strategy that involves matching HES data to community real estate and people registers. “Property registers,” drawn from real estate data, contain the same kind of housing characteristics – namely, the number of beds and baths, total living area, and when the house was built – that the HES data contain. The real estate data additionally include the address of the property and the names of the homeowners. A “people register” is a list of people in a community relevant to the re-identification. For example, a local voter list is a people register of all registered voters in a community and might be useful for re-identifications involving a geographically bounded group of adults. People registers tend to include the names, addresses and demographics of the people in the community. The study was conducted in 2006, so the Attackers sought to construct registers of property as it was in 2006 and of people who lived in these communities in 2006.

The Attackers’ approach unfolds in two phases; see Figure 1. In the first phase, the Attackers match HES data to real estate data on housing characteristics (see the left side of Figure 1). Any matches found will associate named property owners and home addresses to the demographics of participants in the HES who reside at a home having those characteristics. If unique matches result, then re-identifications would likely be done for those HES records. However, houses in these communities tend to have similar number and types of rooms, and most were built at the same time. Therefore, multiple matches are likely for each HES record and so further matching is necessary.

In the second phase, the Attackers match the results from the first stage to a people register on race, gender, age, address, and move-in date, in order to associate names of people residing at those residences. See the right side of Figure 1. Matches associate the name of someone known to live at the address in the people register with an HES participant living in a similar house and having those demographics. Further, if the homeowner name is the same as the name from the match in the people register, and the HES data states the person is a homeowner, then the match is further confirmed.

The Attackers assess the re-identification strategy by identifying risk pools and computing the number of re-identifications for $k<5$, 11, and 20. These re-identifications are believed to contain the correct match, but they do not necessarily contain the correct match. Therefore, each experiment concludes when the Scorers report, by binsize, how many of the proposed groupings include the correct named person or the correct address. Re-identifications of addresses rely on the same matches of resident demographics, but are evaluated separately because addresses themselves are important personal information, as described earlier.

Finally, Attackers also explore variations of the approach based on human matching versus automated matching. Computers can process more records quickly, but humans tend to use heuristics that may provide improved results.

The next subsection provides a stepwise description of the approach for replication and detailed study. The general reader can advance to the Results section without loss of understanding.

Detailed Approach

To test their re-identification strategy, the Attackers acquire and construct appropriate property and people registers and then match records, as described in the 7 steps below.

Step 1. The Scorers construct dataset that satisfies more than the minimum HIPAA Safe Harbor requirements. Starting with the original HES Dataset, the Scorers redact names, addresses, and other personally identifiable information and identifiers (e.g., number of dogs in the home, information on IV treatments, individual room dimensions). All dates (e.g., birth year, year house built, year moved in) are converted to decade and aggregated, so that they are reported in ranges of 10 or more years. Decades were aggregated so that each reported range contained at least 5 records. One field was constructed (total square feet of living area in aggregated ranges) because the Scorers considered individual room dimensions or exact square footage as potentially uniquely identifiable data. We refer to the resulting data as the “HIPAA Dataset”.

Step 2. The Attackers attempt to distinguish Richmond from Bolinas records and, among the Richmond records, Atchison Village from Liberty Village, in order to improve re-identification accuracy. Using information in the HES Publications, the Attackers identify characteristics in the HIPAA Dataset that are specific to homes in Atchison Village, Liberty Village, and Bolinas and then subdivide the records of the HIPAA Dataset into those records likely to be specific to those communities. The Scorers report the accuracy of the Richmond-Bolinas and Liberty Village-Atchison Village subdivisions at the end of the study, even though the Attackers may use these partitions in intermediate steps.

Step 3. The Attackers construct a dataset to use for re-identification. They compute new fields that are convenient for re-identification and eliminate fields that are not relevant to the re-identification strategy. We refer to the result as the “De-ID Dataset.” The De-ID Dataset remains HIPAA Safe Harbor–compliant because it is a subset of the HIPAA Dataset. All re-identification attempts are on the De-ID Dataset.

Step 4. The Attackers construct a property register for the rental units in Liberty Village. The tax assessor data list the Liberty Village complex as one large real estate block. So the Attackers use Google Earth images and rental property websites to infer the addresses, number of baths, and number of bedrooms for each unit. We refer to the result as the “Liberty Village Property Register.”

Step 5. The Attackers construct a property register for Bolinas, California. Unlike the tax assessor data for Atchison Village, the acquired 2006 tax assessor data for Bolinas does not

contain any housing details [38]. Instead, the data for each home include the names of the owners, address, a unique parcel identifier, and various tax values. However, the tax assessor additionally hosts a website on which searches by a parcel identifier yield detailed housing characteristics, such as the number of rooms, baths, and bedrooms for the parcel [39]. The Attackers use the parcel identifiers from the acquired property tax data to construct a "Bolinas Property Register" with the same fields as the Atchison Village Property Register.

Step 6. The Attackers construct registers of people known to have lived in Atchison Village, Liberty Village, and Bolinas in 2006. Many HES participants in Atchison Village and Bolinas are homeowners, but reliance solely on the names found in the property registers may be misleading and limiting, so the Attackers construct registers of people known to be associated with the addresses in these communities. Using information from a data broker, the Attackers search the addresses from the Atchison Village Property Register to identify named people who lived at an Atchison address in 2006, from the Bolinas Property Register to identify those who lived at a Bolinas address in 2006, and from the Liberty Village Property Register to identify those who lived in a Liberty Village unit in 2006. We refer to the resulting registers as the Atchison People Register, the Bolinas People Register, and the Liberty People Register, respectively.

Step 7. The Attackers execute re-identification strategies, and the Scorers report results. There are four sub-steps.

In Step 7a, the Attackers associate records in the De-ID Dataset with known addresses by matching housing characteristics, such as number of baths and bedrooms, in the De-ID Dataset to those in the property registers. Rather than matching against all records in the De-ID Dataset, the Attackers use the partitions derived in Step 3 to match those records in the De-ID considered most relevant to a community. The result is an association of named locations to specific records in the De-ID Dataset.

In Step 7b, the Attackers put names from the people registers to specific records in the De-ID Dataset by matching the combined property and addresses linkages from the results of Step 7a to records in the people register on personal demographics, such as age, gender, and race/ethnicity. The Attackers visually determine gender from the person's first name and Hispanic ethnicity from the person's last name and perform matching manually using a spreadsheet program. The result is an association of named people and locations to specific records in the De-ID Dataset.

In Step 7c, the Attackers repeat Step 7b using a computer program to associate race and gender to last and first names based on statistical occurrences of those names in U.S. Census data and to match records automatically based on personal demographics, such as age, gender, and race/ethnicity. The result is another association of named people and locations to specific records in the De-ID Dataset.

Finally, in Step 7d, the Scorers report on the correctness of the associations (or matchups) separately by community. Scorers report the number of HES participants found in each people register and the number of addresses of HES participants found in each address register. Matchups (or re-identifications) of one or more named people or named locations to

a specific study record are given to the Scorers, who report the number correct per binsize group.

When scoring results, the Scorers apply the following rules:

- Names must match exactly, except in the following cases:
  - Shortened versions of names (e.g., Jon for Jonathan)
  - Commonly accepted nicknames (e.g., Bill for William)
  - Hyphenated last names, where at least one name overlaps (e.g., Jon Smith and Jon Smith-Jones will match)
  - Participant last name is listed as a middle name with a different last name (e.g., maiden name adopted as middle name following marriage)
  - Excepting the above case, middle names and initials will not be considered for matching. Note: if the Attackers re-identified “Katherine Jones” as an HES subject, but the HES Dataset listed “Katherine Smith,” the Scorers would not consider the Attackers successful. However, the scorers would accept Katherine Smith Jones as a match to HES participant Katherine Smith.
  - Obvious misspellings, including non-alphanumeric characters, spacing, and capitalization
- Addresses much match exactly, except in the following cases:
  - Street suffix abbreviations (e.g., St for Street)
  - Street suffix omissions
  - Prefixes used to designate unit may differ, but unit number must match (e.g., Unit 1 and Apt 1 will be accepted as a match, but Unit 1 and Apt 2, or Unit 1 and Unit 2, will not)
  - Word order changes
  - Obvious misspellings

Step 8. The Scorers construct a dataset with exact birth years, which satisfies the minimum HIPAA Safe Harbor requirements. A second version of the HIPAA Dataset, provided after re-identification on the first version, included exact birth year. We refer to this as the “HIPAA Exact Dataset”. The Attackers then repeat the relevant parts of Step 7 using the HIPAA Exact Dataset.

Results

This section walks through the work performed. In the first subsections, the Attackers establish a dataset redacted beyond the HIPAA Safe Harbor standard that provides the basis for re-identification. The Attackers also explain the means used for distinguishing between Atchison Village, Liberty Village, and Bolinas records in the dataset.

The next consecutive subsections report on the construction of property registers for Atchison Village, Liberty Village, and Bolinas. Afterwards, subsections detail the assembly of people registers for each community and report on demographic statistics for each population.

The remaining subsections report on matches of records in the dataset to people and addresses in the registers made after assembly of all the components – the dataset, the property registers and the population registers – and itemize which matches were correct. These results appear in consecutive subsections, one for each of the communities, Atchison Village, Liberty Village, and then Bolinas.

The final subsection repeats the matching experiment having year of birth information in the records of the dataset. The section ends with a comparison of results between data redacted beyond the HIPAA standard to data redacted at a level permissible by the HIPAA standard. The paper ends in the following section with a discussion of the findings.

Results for Step 1. HIPAA Dataset

The Scorers produced the “HIPAA Dataset” from the original HES Study Dataset that goes beyond the minimum HIPAA Safe Harbor requirements. The HES dataset consists of three files, divided into survey data, air measurements, and dust measurements. Appendix A provides a complete list of field descriptions for the files.

The Survey file contained 50 rows, one row for each house sampled. There were 256 fields, including demographic data about the research participants such as race, gender, birth decade group, education level, decade group participant moved into the residence, whether participant owns the home, square footage of living area, number and types of rooms, decade group house was built, and details about the home and the use of various appliances and pesticides.

Dates in the Survey file appeared in decade groups of at least 10 years. Specifically, values for birth were: 1920-1939, 1940-1949, 1950-69, or 1970-1989. Values for move-in date were: 1970-1989, 1990-1999, or 2000-2009. Values for house built date were: 1840-1949 or 1950-1989.

Values for total square footage were: 450-500, 500-650, 650-700, 700-1000, or 1000-2000.

There was no ZIP (or postal code) or other explicit geographical designation in the Survey file.

The Air and Dust files described the compounds found. The Air file had 12,767 indoor and outdoor measurements, and the dust file had 3,871 measurements for the 50 homes.

Results for Step 2. Records for each neighborhood

The Attackers reviewed HES Publications, found the following description of participant demographics, and reviewed online information about the communities.

The study was done 10 years earlier in 2006. Table S1, available as a supplement to the online version of the article [1], provides the following demographic summary.

Participants were 85 percent female and 15 percent male from Richmond and 60 percent female and 40 percent male from Bolinas. In Richmond, 5 percent were less than 26 years in age, 15 were between 26 to 40 years, 43 percent were between 41 and 60 years, and 37 percent were more than 60 years in age. In Bolinas, 10 percent were less than 26 years in age, 20 percent were between 41 and 60 years, and 37 percent were more than 60 years in age.

In Richmond, 41 percent of the participants self-identified as Hispanic, 54 percent self-identified as White, and 11 percent selected another race/ethnicity (3 percent Black, 5 percent Native American, and 3 percent Asian). Participants could self-identify as more than one race. Sixty-two percent were interviewed in English and 38 percent in Spanish. In Bolinas, none of the participants were Hispanic, 89 percent were White, and 44 percent selected another race/ethnicity (11 percent Black, 22 percent Native American, and 11 percent Asian); all were interviewed in English. The racial composition of Bolinas reflects a correction provided by the Scorers due to one person missing race information in Bolinas that was not noted in the original Table S1 [1].

Highest educational attainment in Richmond was as follows: 37 percent had a college education or higher, 26 percent had some college or post-high school training, 5 percent were high school graduates, and 32 percent had completed 11^{th} grade or less. In comparison, 100 percent of Bolinas participants had at least a college degree.

Finally, in Richmond, 79 percent were homeowners compared to 70 percent in Bolinas.

From these characteristics, the Attackers computed the following invariants about the 10 records of Bolinas participants:

- All 10 Bolinas participants spoke English
- 4 were male, 6 were female
- 0 Hispanic, 1 Asian, 1 Black, 2 Native American
- More than one race per person reported
- 3 were renters, 7 were homeowners
- All 10 have a college education or better
- Year of birth groups:
  - 1 was born 1970-1989
  - 2 or fewer were born 1950-1969
  - 9 or fewer were born 1940-1049
  - 7 or fewer were born 1920-1939

None of the 50 records indicated a garage. Forty of the properties were built between 1840 and 1949, five between 1950 and 1989, and five were missing built year.

Based on these findings, the Attackers sought to identify which 40 of the 50 records in the De-ID Dataset belonged to Richmond participants and which 10 records belonged to Bolinas participants by using published information about the study and values that appeared in the 50 records. The three geographical areas had a combined population of about 3,000 adults at the time of the study.

The attackers wrote a computer program to search all possible combinations of 10 of the 50 records that satisfied the demographic constraints. The computer identified 3 combinations of 12 records that satisfied all the demographic constraints for Bolinas. The remaining records would be the 40 for Richmond.

Another published table (reprinted in Appendix B) showed differences between outdoor air samples for homes in Bolinas and those in Richmond [17]. Attackers reviewed the publication for any chemical differences that might distinguish Bolinas and Richmond homes and learned that fluoranthene values for 33 Richmond homes reportedly ranged from 0.41 to 2.7 ng/m^3. For Bolinas, 8 homes did not have detectable levels of fluoranthene, one home had the maximum of 3.8 ng/m3, and one home had an unknown level.

The Attackers manually examined the outdoor fluoranthene measures for the 43 homes and found that there were exactly 33 records in the 0.41 to 2.7 ng/m^3 range and 10 others that conformed to the summary statistics for Bolinas. Therefore, the Attackers believed the 10 records to be the Bolinas records and all others to be Richmond (including those for which no outdoor fluoranthene measurement was available). This configuration also agreed with one of the combinations found by the computer, which further supported the finding.

The Attackers submitted 40 PrivacyIDs as belonging to Richmond participants and 10 PrivacyIDs as belonging to Bolinas participants. The Scorers reported (after the experiment concluded) that the record designations were 100 percent correct. Therefore, the Attackers were able to use previously published results from the study to identify which records belonged to which community (Bolinas or Richmond).

The researchers reported that 40 of the records came from Richmond but did not distinguish how the 40 records split between Atchison Village and Liberty Village. The Attackers examined the 40 Richmond records and found 8 were for renters, 31 were for homeowners and one was unknown. Liberty Village is a rental complex, so all Liberty Village participants should be renters. Atchison Village is a housing cooperative. A “homeowner” in Atchison Village owns a share of the cooperative, and the cooperative decides who lives where. Cooperatives often have specific rules that impose limitations on renting. Therefore, the Attackers concluded that the 31 homeowners were from Atchison Village, the 8 renters were from Liberty Village, and the one unknown could belong to either neighborhood. In summary, the Attackers split the 40 Richmond records into 32 records for Atchison Village and 9 records for Liberty Village, a total of 41 records because one record appears in both groups. The Attackers then used these groupings for re-identifications involving Atchison Village and Liberty Village.

The Scorers reported (after the experiment concluded) that the 32 records for Atchison Village were correct and 8 of the 9 records for Liberty Village were correct. Therefore, the Attackers were able to reasonably ascertain which records belonged to Atchison Village and which to Liberty Village.

Results for Step 3. De-ID Dataset

Based on the Attacker's approach (see Figure 1), the data observations noted, and derivations above, the Attackers identified 15 fields in the HIPAA Dataset and 3 computed fields (the number of rooms and the numbers of bedrooms and bathrooms) as the subject of re-identification. The result is the De-ID Dataset, which contains 50 data rows and 18 fields. The fields include participants' race, gender, decade group of birth, education level, decade group for when they moved into their study residence, and whether they owned the home. Information about a residence includes square footage, room counts, and multi-decade grouping in which the house was built (i.e., 1840-1949 and 1950-1989). Fluoranthene level in outdoor air was also included. Figure 2 provides a summary of the fields in the De-ID Dataset.

Results for Step 4. Atchison Village and Liberty Village Property Registers

Atchison Village Property Register—As described earlier, the Attackers purchased the 2006 tax assessor data for Atchison Village from the County of Contra Costa Assessor's Office [29]. The Atchison Village Property Register had 124 fields and 450 data rows. The fields included the names of the owners, the address of the property, the numbers of bedrooms and baths, the total number of rooms, the year the house was built, and the total living area. The Attackers identified these 8 (of the 124 fields) fields as being relevant to re-identification; see Figure 3a.

All 450 properties had one living unit built in 1942. Most had two bedrooms, one bathroom, and a total of 4 rooms in a living area of 781 square feet (179 of 450 houses or 40 percent). Figure 3b displays the counts of bedrooms, bathrooms, total rooms, and living area in combination.

Because the tax assessor data should be a complete record of all properties in Atchison Village, a list of the addresses from the Atchison Village Property Register should contain all the addresses of HES participants from Atchison Village. So the Attackers submitted the list of 450 addresses. At the end of the experiment, the Scorers reported that 32 of the addresses from HES participants appeared on the list, further suggesting that the remaining eight Richmond addresses are from Liberty Village.

Liberty Village Property Register—At the time of these experiments, Liberty Village was a 100-unit rental complex consisting of 50 duplexes comprised of one-, two- and three-bedroom single-story units that ranged from 528 square feet to 816 square feet [40]. The units were grouped into courtyards with a front and back yard for each residence, were carpeted, and had gas stoves and heating. There was also a clubhouse and a swimming pool. The one-bedroom units were 528 square feet, the two bedroom units were 624 square feet,

and the three bedroom units were 816 square feet [41]. All units had one kitchen, living room, and bathroom. The complex was built in 1942.

The Attackers used aerial Google images [34] to determine the bedroom count for each unit as follows. First, the Attackers measured the length of each roofline to determine the number of bedrooms in a unit. Then, they used parking spot numbers to estimate approximate street numbers for the addresses. Finally, the Attackers associated the addresses with the number of bedrooms for each unit. Below is a walk-through of the approach.

Figure 4 shows a Google image of Liberty Village. The buildings with the brown roofs are Liberty Village. From the image, each building has two pathways leading to the building, implying each building houses two rental units, with the possible exception of the clubhouse area.

The Attackers measured the lengths of the rooflines in a printed image and found that each roof was one of three measurements: 1.9 cm, 2.1 cm, or 2.6 cm. The roofline lengths and the fact that each building had two pathways led to the following inferences: each of the smallest buildings houses two one-bedroom units; each of the middle-sized building houses two of the two-bedroom units; and, each of the largest buildings houses two of the three-bedroom units. The 4 small buildings (red lines in Figure 5) identified 8 one-bedroom units. The 34 middle-sized buildings (green lines in Figure 5) identified 64 two-bedroom units. Finally, the 12 large buildings (blue lines in Figure 5) identified 24 three-bedroom units. In total, the Attackers graphically identified 100 units, which was the correct total number of units in Liberty Village.

The aerial Google images name streets. The Attackers used parking spot numbers to infer the street numbers of the units, as practical (yellow circles in Figure 6). The Google Earth interface allowed a user to identify some addresses by hovering over rooftops. The Attackers refined these addresses based on the parking spaces and unit address patterns (rectangles in Figure 6). The result was 111 addresses for the 100 units because of ambiguity with some addresses. The final street addresses were from 7 to 168 Chanslor Circle, from 6 to 30 Chanslor Row, from 14 to 24 Circle Court, from 118 to 217 Chanslor Avenue, and from 115 to 348 West Chanslor Avenue.

The 111 addresses identified as being in Liberty Village had 3 configurations based on the number of bedrooms. The most commonly occurring home had two bedrooms, one bathroom, a living room, and a kitchen, for a total of 6 rooms in a living area of 624 square feet. Figure 7 lists the counts of the three configurations.

The final result was the Liberty Village Property Register, which listed the address, number of bedrooms and baths, total rooms, year house built, and square footage for each of the 100 Liberty Village units at 111 addresses. See Figure 7.

The Attackers sent a list of the 111 addresses from the Liberty Village Property Register to the Scorers. The Scorers reported, at the end of the experiment, that 5 of the addresses from HES participants appeared on the list. Together with the Atchison Property Register, the

Attackers identified 37 out of 40 Richmond addresses. The HES researchers never disclosed the number of Liberty Village versus Atchison participants in their publication.

Results for Step 5. Bolinas Property Register

As described earlier, the Attackers purchased the 2006 Bolinas Tax Data [38]. Unlike the tax data for Atchison Village, the Bolinas tax data did not contain housing characteristics; specifically, it did not contain the number of rooms, bedrooms, baths, or total living area. Instead, the tax data for Bolinas included the address and owners of the property, the number of units on the property used for living, and then various fields related to the tax computations, such as land value. There were 26 fields and 1,583 data records. However, only 626 of the data records were properties that had units for living; the other records concerned land that apparently had no property on which people lived. Of the 626 real estate properties where people lived, only 610 had addresses listed. Most of the 610 addresses had single-family dwellings (572 of 610 or 94 percent). The median and average were homes with one living unit, and the standard deviation was 0.4. One property had the maximum of 6 units in which people lived.

The Attackers constructed a table with these fields: property id, the number of units on the property in which people live, the names of the owners, and the address for each property that had living units. We term this the “Bolinas Tax Data table.”

The Marin County Tax Assessor’s office had a website that displayed the number of bedrooms and baths and other housing characteristics for a property once a “property id” is given [39, 42]. Figure 9 steps through the pages of the website to display housing characteristics for the randomly selected residence having property id 188-100-05.

Using the website added uncertainty, because the identity of the homeowners was from 2006 tax data but the housing characteristics were mined from the website in 2013 (and replicated in 2017). An HES participant from Bolinas could have made home renovations that changed the number of bedrooms or bathrooms during this time, and if so, the recorded information would not match the 2006 information.

Regardless, the Attackers automated the process shown in Figure 8 by writing a Python program that used the property (parcel) ids from the Bolinas Tax Data to walk through the website in the same way a human would to retrieve the housing characteristics for each of the 610 Bolinas residences. If an error was encountered, the Attackers then searched for properties having the same first groups of digits on the parcel number and the same owner that did not otherwise appear on the list. In these cases, the parcel numbers may have changed between the date of the tax data (2006) and the web searches of housing characteristics (2013 and replicated in 2107), so this was a means of locating the new property id to fetch the housing characteristics. The website provided housing characteristics for 533 of the 610 parcels. Searches for the remaining 77 parcels gave an error, and no other parcel id was found for the property.

Of the 533 parcels found on the assessor’s website, 3 were built after 2006, so they were dropped. The final result was housing, address, and ownership information for 530 of 610

(or 87 percent) of the Bolinas residential properties. Of these 530 addresses, 105 (20 percent) had parcel changes since 2006, most of which were subdivisions. An unknown number of parcels may have further changed, and likely increased, the number of bedrooms or baths since 2006.

The Attackers constructed a file that associated the owner names and property addresses from the Bolinas Tax Data with the housing characteristics retrieved from the website for the same property; this is the “Bolinas Property Register.” It had 8 fields and 530 data rows. Figure 11 lists its fields.

The Bolinas Property Register is similar to the Atchison and Liberty property registers, except it does not include the total number of rooms. It does include the names of owners, whereas the Liberty Property Register includes no names. The Bolinas Property Register additionally includes the square footage of the garage.

The Bolinas properties were far less homogeneous than the Atchison Village and Liberty Village properties. Of the 530 properties in the Bolinas Property Register, 477 (or 90 percent) had a unique combination of bedrooms, baths, and living area, with the variability being greatest in the amount of living area. For example, the largest number of residences having the same combination of bedrooms, baths, and living area was 7 for homes having 2 bedrooms and one bathroom and 768 square feet. Even though 158 (or 30 percent) of the 530 homes had 2 bedrooms and one bathroom, the possible living areas ranged from 465 to 2,338 square feet, with a median of 968 square feet, an average of 1,057 and a standard deviation of 361. Figure 10 shows descriptive statistics for each housing characteristic separately.

The Attackers sent a list of the 530 addresses that constitute the Bolinas Property Register to the Scorers. The Scorers reported, at the end of the experiment, that 9 of the 10 Bolinas addresses of HES participants (or 90 percent) appeared in the Bolinas Property Register.

Results for Step 6. People Registers

At this point, the Attackers had constructed three property registers, one each for Atchison Village, Liberty Village, and Bolinas. Later, the Attackers used these property registers in the first stage of the re-identification, as depicted on the left side of Figure 1. The second stage of the re-identification required the construction of people registers, which are lists of named people known to have lived at the addresses during the study period.

The names of homeowners from the Atchison Village property data were not used as a people register because the Atchison Village Cooperative accommodates relocations within Atchison Village. As a consequence, the tax data for an individual property may not reflect the true resident at the time of the study. The names of homeowners from the Bolinas property data were not used as a people register because some participants from Bolinas rented. So the Attackers constructed people registers for all 3 communities using the following 3 steps:

1. Start with a blank people register. The fields are: address, name, birth year, move-in and move-out years, gender, and race.

2. For each address in the property register:

   Search a public data broker's website [31, 32] for people who lived at the address during 2006 (the study period).

   a. For each person found:

      Add a record to the people register that contains the person's name, birth year, the earliest year they were known to live at the address, and the year they moved out, if provided.

3. For each name acquired in (2) above:

   a. Infer the person's gender, as possible, from the person's first name and append the information to the person's record in the people register.

   b. Infer the race or ethnicity of the person, as possible, from the person's last name (also known as the family or surname) and append the information to the person's record in the people register. Using last names to infer race is not a good predictor of Blacks because Whites and Blacks often share a last name. Therefore, some number of those identified as white may be black using last name inference.

The subsections below describe the demographics of the people register in more detail than is necessary to interpret the results. The reader can advance to the summary subsection, Summary of Property and People Registers, without loss of information sufficient to understand the results. Meanwhile, the reader seeking a deeper understanding of the communities involved in this study should proceed.

Atchison Village People Register—The Attackers searched each of the 450 addresses from the Atchison Village Property Register on the public data broker's website. Names and demographics for 1,290 adult residents were found for 434 (or 96 percent) of the addresses; 16 addresses reported no residents.

Almost half the addresses (213 of 434 or 49 percent) had 1 or 2 adult residents. Figure 12(a) shows the distribution of the number of adults per address: 1,127 of 1,290 (or 87 percent) of the people had a year of birth. Figure 12(b) shows the distribution of the birth years of the adults. The youngest people were born in 1993 and the oldest in 1900. The median year of birth was 1956 and the average was 1954 with a standard deviation of 19 for the 1,127 people having birth year information.

Figure 12(c) shows the distribution of move-in dates. The person living at their Atchison Village address the longest moved into the residence in 1970. The median year in which people moved to their Atchison address was 1998, and the average was 1997 with a standard deviation of 7 for the 1290 people.

Many people (142 of 1,290 or 11 percent) moved into their Atchison residence during the year of the study, while 76 residents moved out during the study year of 2006.

The Attackers wrote a Python program that used a list of the 1,645 most popular first names and their frequency of gender usage in the 1990 U.S. Census [35] to assign gender. Of the 1,290 names for residents identified in Atchison Village, 560 (43 percent) had first names more likely to be associated with males, and 582 (45 percent) had first names more likely to be associated with females. Gender was assigned to 1,142 of 1,290 (89 percent) of the names in the Atchison Village People Register.

The Attackers wrote a Python program that used a list of the 151,671 most popular last names and their frequency by race and ethnicity in the 2000 U.S. Census [36]. Of the 1,290 names found for residents in Atchison Village, 594 (46 percent) had last names more often associated with Whites, 533 (41 percent) had names more often associated with Hispanics, and 77 names (6 percent) could not be assigned an inferred race. Figure 12(d) shows the distribution by race and ethnicity. Values in the race field in the Atchison Village People Register were assigned accordingly.

The Attackers submitted the names of the 1,290 residents that they found as residents of Atchison Village in 2006. At the end of the experiment, the Scorers reported that 32 of the names from HES participants appeared on the list of 1,290 residents.

Liberty Village People Register—The Attackers searched each of the 111 addresses from the Liberty Village Property Register on the public data broker's website. Names and demographics for 438 adult residents were found for 98 addresses. Recall, Liberty Village actually had 100 units with 100 addresses, and the Attackers had derived 11 additional addresses. The data broker's website found some addresses defunct, yielding people for 98 of 100 (98 percent) units.

Almost half the addresses (44 of 98 or 45 percent) had 1, 2, or 3 adult residents; see Figure 13(a). Only 303 of 438 (or 69 percent) of the people had year of birth. Figure 13(b) shows the distribution of the birth years. The youngest person was born in 1996 and the oldest in 1900. The median year of birth was 1970, and the average was 1967 with a standard deviation of 14 for the 303 people. Figure 13(c) shows the distribution of move-in dates. The person living at their Liberty Village address the longest moved into the residence in 1980. The median year was 2004, and the average was 2003 with a standard deviation of 4 for the 438 people.

Many people (158 of 438 or 36 percent) moved into their Liberty Village residence during the year of the original study. Many people (112 of 438 or 26 percent) also moved out of their Liberty Village residence during the study year. The number of named residents that the Attackers found who neither moved in nor moved out of Liberty Village during 2006 was 244 (of 438 or 56 percent) of the residents, which is about half of all the Liberty Village residents that the Attackers identified.

The Attackers used their Python program (described earlier) to assign gender to the names of the people identified as living in Liberty Village during the study year. Of the 438 names of residents identified in Liberty Village, 209 (48 percent) had first names more likely to be associated with males, and 161 (37 percent) had first names more likely to be associated

with females. Gender was assigned to 370 of 438 (84 percent) of the names in the Liberty Village People Register.

The Attackers used their Python program (described earlier) to assign race and ethnicity to the names of the residents identified as living in Liberty Village during the study year. Of the 438 names found for residents in Liberty Village, 316 (72 percent) had last names more often associated with Hispanics, 87 (20 percent) had names more often associated with Whites, and 23 (5 percent) had no race or ethnicity assigned. Figure 13(d) shows the distribution. Overall, 415 of 438 (95 percent) of the surnames were assigned a race or ethnicity.

The Attackers submitted the names of the 438 residents that they found as residents of Liberty Village in 2006. At the end of the experiment, the Scorers reported that 3 of the names from HES participants appeared on the list of 438 residents.

Bolinas People Register—The Attackers searched each of the 530 addresses from the Bolinas Property Register on the public data broker's website. Names and demographics for 1,082 adult residents were found for 465 (87 percent) of the addresses; 67 addresses reported no residents.

More than half the addresses (317 of 465 or 68 percent) had 1 or 2 adult residents; see Figure 14(a). Most people, 1,038 of 1,082 (96 percent), had a year of birth; only 44 did not; see Figure 14(b). The youngest person was born in 1992 and the oldest in 1908. The median year was 1952, and the average was 1954 with a standard deviation of 16 for the 1,038 people. Figure 14(c) shows the distribution of move-in dates. The person living at their Bolinas address the longest moved into the residence in 1963. The median year was 1994, and the average was 1994 with a standard deviation of 8 for the 1,082 people.

Some people (83 of 1,082 or 8 percent) moved into their Bolinas residence during the year of the study. Similarly, some Bolinas residents (90) moved out during the study year of 2006. The number of named residents that the Attackers found who neither moved in nor moved out of Bolinas during 2006 was 949 people.

The Attackers used their Python program (described earlier) to assign gender to the names of the people identified as living in Bolinas during the study year. Of the 1,082 names of residents identified in Bolinas, 476 (44 percent) had first names more likely to be associated with males, and 495 (46 percent) had first names more likely to be associated with females. Gender was assigned to 971 of 1,082 (90 percent) of the names in the Bolinas People Register.

The Attackers used their Python program (described earlier) to assign race and ethnicity to the names of the residents identified as living in Bolinas during the study year. Of the 1,082 names found for residents in Bolinas: 52 (5 percent) had last names more often associated with Hispanics, 819 (76 percent) had names more often associated with Whites, and 161 (or 15 percent) had no assignment; see Figure 13(d). Overall, 921 (of 1082 or 85 percent) of the surnames were assigned a race or ethnicity.

The Attackers submitted the names of the 1082 residents that they found as residents of Bolinas in 2006. At the end of the experiment, the Scorers reported that 5 of 10 (or 50 percent) of the names from HES participants appeared on the list of 1,082 residents.

Summary of Property and People Registers—Each property register was produced in a distinct manner. The Atchison Village Property Register, containing 450 addresses, came directly from the tax assessor data, and therefore should contain all addresses in Atchison Village with their appropriate housing characteristics. Liberty Village was a rental community of 100 units. The Attackers constructed a property register by inferring 111 addresses and housing characteristics for the 100 units. The Bolinas Property Register started with tax information to identify addressed parcels having living units, but the Attackers mined the housing characteristics from a tax assessor website almost 10 years after the study for the list of 530 addresses. Therefore, the Atchison Village and Liberty Property Registers appear to be the most complete and Atchison the most accurate.

The communities differed in their property characteristics and homogeneity. Atchison Village and Bolinas were both communities of primarily homeowners, whereas Liberty Village was a rental complex. The housing characteristics – number of bedrooms and baths and living area – of Atchison Village and Liberty Village were homogeneous. Most homes had two bedrooms and one bathroom built in 1942.

Of the 530 properties in the Bolinas Property Register, 477 (90 percent) had a unique combination of beds, baths, and living area with the variability being greatest in the amount of living area. The houses were built between 1879 and 2005 with a median year of 1959. The largest number of residences having the same combination of bedrooms, baths and living area was only 7 homes, which had 2 bedrooms and one bathroom with a living area of 768 square feet.

The Attackers constructed population registers using information available from a data broker. While the information seemed reasonable and comprehensive, there was no guarantee that the data were accurate or complete.

The characteristics of residents differed among the communities. The residents found for Atchison Village had last names the computer program associated with Whites (46 percent) and Hispanics (41 percent). The names found for residents in Liberty Village were much more frequently associated by the computer program with Hispanics (72 percent) than Whites (20 percent). The computer program associated names of Bolinas residents predominantly with Whites (76 percent) with few Hispanics (5 percent).

Liberty Village experienced a lot of mobility during the study year (as reported earlier in Liberty Village People Register). A total of 76 of the 98 units (78 percent) changed occupancy during the year of the study. A third of the properties in Atchison Village changed occupants during the study year (150 of the 450 or 33 percent). However, few of the Bolinas residences changed occupancy during the same year (40 of the 530 or 8 percent of the addresses).

The characteristics described above about these people registers place important limits on re-identification attempts. Here is a summary based on a comparison of demographic homogeneity, the number of adults per residence, mobility, and data quality.

The more homogeneous a community, the more difficult it is to acquire correct small group re-identifications because many different people and homes share the same features indistinguishably. All three groups are homogeneous, but there is a noticeable difference in racial homogeneity. The Liberty Village People Register is almost all Hispanic, and the Bolinas People Register is almost all White. Only the Atchison Village People Register has substantive variability in race (about half Hispanic), so we might expect more matches from Atchison Village than the other two. Liberty Village had the most homogeneous property register.

A population having more adults per residence will likely make larger groups in matches of people by name based on housing. The Liberty Village People Register has many more people on average per household than does Atchison Village or Bolinas, so we might expect Liberty Village to have fewer small group re-identifications.

The greater the number of people moving in and out of a residence during the year of the study, the more difficult it is to match a person to a residence because the register and the data may show different residents for the same year. Liberty Village had the greatest mobility during the study period.

Finally, using Bolinas housing characteristics that are 10 years newer than the study data would be expected to limit correct matchups of Bolinas data.

Results for Step 7. Re-identifications

Matching Characteristics

Matches between the property data and the HES records in the De-ID Dataset should have no garage, the same number of bedrooms and baths, and, in the case of Richmond homes, be built in 1942.

At least 32 of the 50 properties (64 percent) in the De-ID Dataset were from Atchison Village, yet the property square footages did not match those listed in the De-ID Dataset even accounting for smaller units at Liberty Village. The Attackers modeled the records in the De-ID Dataset as a random sample drawn from the three different communities and then compared the distributions of their square footages for living units. The Attackers assumed the distributions in the De-ID dataset were representative of the population, so normalizing the distributions allowed the Attackers to associate values in the De-ID dataset with likely equivalents in the property data based on the following cut-offs used for all three communities.

totalsquareft in De-ID Dataset	Total Living Area in Property Data
400-450	<= 750
450-500	>=700 and <=790
500-650	>=736 and <=890
650-700	>=840 and <=910
700-1,000	>=875 and <=1,125
1,000-2,000	>= 1,080

A record in a people register matches a record in the De-ID Dataset if the person's birth year, move-in date, gender, and race or ethnicity assignment agrees. A person's year of birth would match to one the following ranges: 1920-1939, 1940-1949, 1950-1969, and 1970-1989. Similarly, a person's move-in year in a people register matches one of the following ranges found in records in the De-ID Dataset: 1970-1989, 1990-1999, or 2000-2009. Gender is Male or Female and race is one or more of: Black, White, Hispanic, Asian, or Native American.

Two sets of matches exist based on whether a computer or human assigned gender and race. The computer-assigned set also includes whether matches to missing values for gender or race in the people register are used or not. The set whose values for race and gender were tagged manually matches only to non-blank results. Below we report on the most relevant results.

The subsections below describe details of matchups and the scoring of matchups in detail by community. A summary of results starts the Discussion section.

Atchison Village Re-Identifications

The Attackers matched the records in the Richmond records of the De-ID Dataset to those in the Atchison Village Property Register based on housing characteristics using a computer program written by the Attackers; see Figure 1(a). The program matched records having the same number of bedrooms and baths, having been built in 1942, and having living areas consistent with the ranges described earlier.

The result was 3,813 matches for 32 of 40 (or 80 percent) of the Richmond records. These matches agreed with the 32 records the Attackers considered to be Atchison Village records. The Attackers continued their analysis with these 32 possible Atchison Village records.

Two records in the De-ID Dataset each matched to only one record in the Atchison Village Property Register. In terms of the maximum number of matches, 3 records in the De-ID Dataset matched to 223 different records in the Atchison Village Property Register. The same record in the Atchison Village Property Register may match to more than one record in the De-ID Dataset. Figure 15 shows the cumulative number of Richmond records in the De-ID Dataset matched to records in Atchison Village Property Register by binsize.

This matching result combines the Atchison Village Property Data to the De-ID dataset. Subsequent steps use this combined dataset.

Atchison Village Re-Identifications: Named People with Hand Labels—The Attackers matched the combined Atchison Village property and De-ID data to records in the Atchison Village People Register using a Python program that the Attackers wrote. The program matched records based on birth year and move-in year information; see Figure 1(b). Gender and race were not used. The Attackers then manually assigned race and ethnicity and gender to the matching records and concluded the matching manually using information from HES Publications and a spreadsheet program.

The Attackers found a total of 121 matches of people from the Atchison People Register to records in the combined Atchison Village De-ID and property data for 17 of 32 (53 percent) of the Atchison records having a small group (k<20) re-identification. Five matches were unique re-identifications. For $k<5$, the risk pool was 9 people for 7 re-identifications. For $k<11$, the risk pool was 40 people for 12 re-identifications. For $k<20$ the risk pool was 109 people for 17 re-identifications. Some people appear in more than one group. Figure 16 shows the accounting of binsizes for the matches.

Atchison Village Re-Identifications: Addresses with Hand Labels—The Attackers also examined the re-identification of addresses from the same Atchison Village People Register in which the Attackers manually assigned gender and race. The Attackers found a total of 135 matches of addresses from the Atchison Village People Register to records in the combined Richmond De-ID and Atchison Village property data for 18 of 32 (56 percent) of the Atchison Village records having a small group (k<20) re-identification. As had been the case with matches to named people, 5 were unique re-identifications. For $k<5$, the risk pool was 9 addresses for 7 re-identifications. For $k<11$, the risk pool was 39 addresses for 12 re-identifications. And, for $k<20$ the risk pool was 109 addresses for 18 re-identifications. Some addresses appear in more than one group. Figure 17 shows the accounting of binsizes for the matches.

Atchison Village Re-Identifications: People with Computer-Assigned Labels—The Attackers then matched the combined Atchison Village property and De-ID data to records in the Atchison Village People Register using a Python program that the Attackers wrote that included matches on the computer assignment of values for gender and race. Matches were based on birth year, move-in date, race and ethnicity, and gender. Blank entries were not matched.

The computer program found a total of 162 matches of people from the Atchison Village People Register to records in the combined Atchison Village De-ID and property data for 21 of 32 (66 percent) of the Atchison Village records having small group (k<20) re-identifications. Two matches were unique re-identifications. For $k<5$, the risk pool was 15 people for 7 re-identifications. For $k<11$, the risk pool was 58 people for 14 re-identifications. And, for $k<20$ the risk pool was 124 people for 21 re-identifications. Some people appear in more than one group. Figure 18 shows the accounting of binsizes for these matches.

Atchison Village Re-Identifications: Addresses with Computer Assigned Labels—The Attackers also examined the re-identification of Atchison Village addresses from the same data in which a computer program assigned gender and race. The Attackers found a total of 159 matches of addresses from the Atchison Village People Register to records in the combined Atchison Village De-ID and property data for 21 of 32 (66 percent) of the Atchison Village records having a small group (k<20) re-identification. There were 2 unique re-identifications. For $k<5$, the risk pool was 16 addresses for 7 re-identifications. For $k<11$, the risk pool was 57 addresses for 14 re-identifications. And, for $k<20$ the risk pool was 108 addresses for 21 re-identifications. Figure 19 shows the accounting of binsizes for the matches.

Atchison Village Re-Identifications: Results—The Attackers sent small group results for addresses and named people for Atchison Village re-identifications to the Scorers to identify which, if any, of the groups sized k<20 had a correct match. Figure 20 provides a summary of the results sent to the Scorers. Of the hand-labeled groups, 7 of the 17 (41 percent) named people groups contained the correct person, and 10 of 18 (56 percent) address groups contained the correct address. The computer labeled matches scored better. One of the 2 unique re-identifications was correct for the named person and the address, 16 of the 21 (76 percent) of the named people groups k<20 contained the correct person, and 16 of 21 (76 percent) of the address groups k<20 contained the correct address. Figure 21 and Figure 22 show the detailed results.

Overall, the Attackers correctly identified the 32 records from Atchison Village and correctly and uniquely identified 1 of 32 (3 percent) by name and address.

Liberty Village Re-Identifications

The Attackers associated 9 records in the Richmond subset of the De-ID Dataset to Liberty Village (8 renters and 1 whose home ownership was not known), and then matched those 9 records to records in the Liberty Village Property Register based on housing characteristics using a computer program written by the Attackers; see Figure 1(a). Matches had the same number of bedrooms and baths, were built in 1942, and had living areas consistent with the ranges described earlier.

The result was 623 matches for the 9 records. None of the matches was unique, and there were no small group matches, a result reflecting the homogeneity of the rental units. Figure 23 (left) lists the cumulative number of records in the Liberty Village Property Register matching to the Richmond records in the De-ID Dataset. Figure 23 (right) shows the cumulative number of Richmond records known to be renters in the De-ID Dataset matched to records in the Liberty Village Property Register by binsize. Subsequent steps use this combined dataset.

Liberty Village Re-Identifications: Hand Labels—Just as was done with Atchison Village, the Attackers matched the combined Liberty Village property and De-ID data to records in the Liberty Village People Register using a Python program that the Attackers wrote, notwithstanding concerns mentioned earlier – i.e., homogeneity of the population and property, and the greater number of residents per household. Gender and race were not used.

The Attackers then manually assigned race/ethnicity and gender to the matching records and concluded the matching manually using a spreadsheet program.

The Attackers found a total of 76 matches of people from the Liberty People Register to records in the combined De-ID and property data. The matches were for 7 of the 9 records that the Attackers believed were from Liberty Village. For *k=1* and *k<5*, the risk pool was 1 person for 1 re-identification. For *k<11*, the risk pool was 8 people for 4 re-identifications. And, for *k<20* the risk pool was 26 people for 7 re-identifications.

Similarly, the Attackers found a total of 70 matches of addresses from the Liberty Village People Register to records in the combined Richmond De-ID and Liberty Village property data. The matches were for 7 of the records believed to be from Liberty Village. For *k=1* and *k<5*, the risk pool was 1 person for 1 re-identification. For *k<11*, the risk pool was 7 addresses for 4 re-identifications. And, for *k<20* the risk pool was 24 addresses for 7 re-identifications.

Liberty Village Re-Identifications: Computer Assigned Labels—Just as was done with Atchison Village, the Attackers then matched the combined Liberty Village property and Richmond De-ID data to records in the Liberty Village People Register using a Python program that the Attackers wrote that included matches on the computer assignment of values for gender and race. Matches were based on birth year, move-in date, race and ethnicity, and gender.

When matches were restricted to records in the Liberty Village People Register that had complete values – i.e., no missing move-in or birth year information, no small group re-identifications resulted. Therefore, computer matches were insufficient with complete information and possible but speculative with missing information.

Liberty Village Re-Identifications: Results—The Attackers sent the hand-labeled results for addresses and named people for Liberty Village re-identifications to the Scorers to identify which, if any, of the groups had a correct match. The Scorers reported that 1 of 7 named people groups contained the correct person and 2 of 7 address groups contained the correct address (Figure 24).

Overall, the Attackers believed 9 records were from Liberty Village, and 8 of those records were actually from Liberty Village. Regardless, the Attackers did not correctly identify any small group re-identifications ($k<5$).

Bolinas Re-Identifications

Just as was done with the other communities, the Attackers matched the records in the Bolinas records of the De-ID Dataset to those in the Bolinas Property Register based on housing characteristics using a computer program written by the Attackers; see Figure 1(a). The program matched records having the same number of bedrooms and baths, being built in the same time period, and having living areas consistent with the ranges described earlier.

The result was 200 matches for 10 of 10 (100 percent) of the Bolinas records. None of the records in the De-ID Dataset matched uniquely. Figure 25 shows the cumulative number of Bolinas records in the De-ID Dataset matched to records in Bolinas Property Register by binsize. This curve is more similar to that for Atchison Village (Figure 18) than for Liberty Village (Figure 23), suggesting that there exists sufficient variability for matching. However, concerns about the quality of the matches remain because of the property data are from 2017 and the HES data are from 2006.

The Bolinas registers did not lend themselves to reliable manual matches. While some derivations were possible, the Attackers lacked confidence in the results because of the homogeneity of the people, small sampling fraction, and the lack of reliability in the housing data. Similar limitations existed for the matches using the computer-labeled data. For $k<20$, the risk pool was 18 people and addresses for 7 re-identifications. As we anticipated might be the case, none of the re-identifications was correct upon scoring.

At the HIPAA Safe Harbor

We repeated the experiments again using HES data that had exact year of birth (rather than decadized), which is permitted by HIPAA. Rather than reporting a study participant's age in bands of 10 or 20 years, we produced re-identifications using a De-ID Dataset that was the same as previously described except that the year of birth was provided. Information about the move-in date and when the house was built remained grouped in ranges of 10 or 20 years.

We matched records in the Atchison Village People Register to the combined Atchison Village property and De-ID records using a Python program that the Attackers wrote. Matches were based on birth year, move-in date, race and ethnicity, and gender. Blank entries were not matched. The computer program found 11 unique re-identifications for named people and addresses. For $k<5$, $k<11$ and $k<20$, the risk pool was the same 27 named people and addresses for 18 re-identifications. Figure 26 shows the accounting of binsizes for the matches.

The Attackers sent these results to the Scorers, who reported that 8 of the 11 (73 percent) unique re-identifications of named people were correct, and 9 of the 11 (82 percent) unique re-identifications of addresses were correct. For $k<5$ (same as for $k<20$), 15 of the 18 (83 percent) groups had a correct named person, and 16 of the 18 (89 percent) groups had a correct address. The Attackers uniquely ($k=1$) and correctly identified 8 of 32 (25 percent) Atchison Village records by name and 9 of 32 (28 percent) by address in the version of the HIPAA-compliant data having exact year of birth, compared to 1 of 32 (3 percent) by name or address in the version of the HIPAA-compliant data having year of birth grouped in 10- or 20 year ranges and using computer-assigned labels.

Similarly, for binsizes $k<5$, the Attackers correctly identified 15 of 32 (47 percent) Atchison Village records by name and 16 of 32 (50 percent) by address in the version of the HIPAA-compliant data having exact year of birth. The correct identifications dropped to 4 of 32 (13 percent) by name or address in data having year of birth grouped in 10- or 20-year ranges and using computer-assigned labels.

Even with the actual year of birth, no reliable re-identifications resulted for Liberty Village or Bolinas.

Discussion

We evaluated the potential for re-identifying 10 people and addresses from Bolinas, and 40 people and addresses from two communities in Richmond (32 from Atchison Village and 8 from Liberty Village), California, by matching a dataset redacted at and beyond the HIPAA Safe Harbor standard with constructed people and property registers. Our results demonstrate that high rates of re-identification are sometimes possible even with heavily redacted data. When less-redacted data (including exact birth year) from Richmond were matched to Atchison Village registers with computer-inferred values for race and gender, we uniquely and correctly identified 8 of 32 (25 percent) by name and 9 of 32 (28 percent) by address. With year of birth grouped in periods of 10 or 20 years, we uniquely and correctly identified 1 of 32 (3 percent) by name and address, 4 of 32 (13 percent) as being one of fewer than 5 names or addresses, and 16 of 32 (50 percent) as being one of fewer than 20 names or addresses.

Anecdotal wisdom suggests that re-identification experiments on heavily redacted data should fail, and this was somewhat true for people and addresses in Liberty Village and completely true for Bolinas. The differences in re-identification rates between these three communities reflect differences in the demographic makeup of the communities and the quality and availability of property data.

All three communities had populations that were fairly homogeneous with respect to age and gender. Liberty Village was predominantly Hispanic and Bolinas predominantly white. Both Atchison Village and Liberty Village were constructed in 1942 and are comprised of a few types of living units repeated many times, but homes in Atchison Village have substantially greater variation in room count and living area. Additionally, accurate property data were available for purchase in Atchison Village, whereas property data had to be inferred for Liberty Village, which is a rental complex. Liberty Village, as a rental complex, also had much higher rates of mobility during the study year, thereby increasing the pool of possible adults living in each home compared to Atchison. These factors likely contributed to the higher rates of re-identification in Atchison Village compared to Liberty Village.

Bolinas differed from both Richmond communities in several important ways. Unlike Liberty and Atchison Villages, Bolinas is not a housing development, so there is much greater variation among homes. However, the quality of the housing data, obtained 10 years after the study with 20 percent of the records having parcel changes, potentially diminished re-identification capability. Bolinas, like Liberty Village, is substantially more racially homogeneous than is Atchison Village, and the sampling fraction was lower in Bolinas (approximately 0.6 percent versus 2 percent). These factors may have contributed to the lower rate of re-identification observed in Bolinas.

These findings suggest that there is something fundamentally flawed with ad hoc redactions of data. They fail to accurately account for the quality and nature of external information. Heavily redacted data may look anonymous, but it is not necessarily so.

The number of correct re-identifications found in the HIPAA Safe Harbor–compliant data having exact year of birth is remarkable (25 percent uniquely and correctly re-identified by name). HIPAA Safe Harbor does not purport to render data with no risk of re-identification.

Instead, the wording states that there exists a “minimal risk,” but the regulation itself does not directly define what a minimal level of risk may be. Prior studies found far fewer re-identifications, less than 0.05 percent of unique re-identifications [2, 3], suggesting that the notion of minimal risk defined by the HIPAA Safe Harbor was that low. Correctly and uniquely re-identifying 25 percent of the people and 28 percent of the addresses is a substantial increase in demonstrated vulnerability.

Earlier studies about vulnerabilities in HIPAA Safe Harbor data narrowly relied on demographic fields (e.g., year of birth, gender and the first 3 digits of ZIP) as the basis for matching. Our study used those and other, non-demographic fields, such as the number of rooms and baths, to link to other data sources. These novel linkages increased the rate of re-identifications.

In fact, critical to our re-identifications was the use of property and people registers. The existence of registers is not unique to household studies. The preamble to the HIPAA Privacy Rule makes reference to Sweeney’s earliest re-identification in which hospital data were matched to a voter list registry to re-identify the medical record of William Weld, then Governor of Massachusetts [44]. Sweeney’s focus on demographics in that seminal example led to a focus on demographic fields in the HIPAA Privacy Rule itself. However, direct matching on demographic fields is not the only vulnerability. Instead, a series of registries can be used to link to other fields in the HIPAA Safe Harbor data before making the final link to a register of named individuals. In our study, the de-identified data were linked to real estate tax data to learn candidate addresses and then matched to the demographics of people associated with those addresses. While tax registries are not generally applicable to medical data, comparable registries in the medical data context today include prescriptions and disease-specific marketing data [45].

Extending HIPAA from healthcare to property data, as we did in these experiments, allowed us to match the HIPAA-compliant study data to identifiable property data using property fields that were present in both data. If all health data were covered by HIPAA, then it would be reasonable to believe that any other dataset containing the same medical fields would also be covered by HIPAA and therefore re-identification attempts would only be able to match medical fields to datasets that had the same redacted demographics, and no names or addresses. However, not all health data is covered by HIPAA, so following this same approach it is possible to link HIPAA-compliant health data with identifiable health data using medical fields. In prior work, Sweeney et al. surveyed flows of health data and found that about half of the more than 2,000 flows of health data they documented were not covered by HIPAA [46]. Among these, they found that 33 states collected and shared

hospital discharge data publicly. Because these statewide datasets are not covered by HIPAA, 30 of the publicly available versions used standards weaker than HIPAA for redaction. In other work, Sweeney correctly re-identified records in one state's hospital discharge dataset by using details from newspaper stories to associate names to records [47]. Once re-identified, the records could theoretically be further matched to HIPAA-compliant data on medical and demographic fields, and thus be used to re-identify the HIPAA-compliant data.

There are many other forms of health data that could also be used to re-identify HIPAA compliant health data. Disease specific marketing lists, for example, include patient names, addresses, and diseases [45]. In a survey of mobile apps, Zang et al. found that personal health monitoring and assistance apps often collect disease specifics, including the date of onset and severity of symptoms, along with the person's name and phone number [48]. A medication refill reminder app contains medication information, from which diagnoses—and even severity of the disease—can be inferred. Datasets of subscribers to health websites and disease discussion lists may include disease information along with names and email addresses. There are many possible sources of health data that include names and contact information as well as medical information, and any of these can be used to re-identify related HIPAA-compliant data.

Our study used three geographical areas whose total population was about 3,000 adults. The HIPAA prescription for reporting geography is the same for data having 3,000 people as it is for populations having up to 20,000 people. No ZIP codes appear. Still, knowing that the data came from three known communities narrowed consideration to a 2 percent sample drawn from 3,000 people.

More generally, there may be a false belief that the HIPAA Safe Harbor only applies to large, datatsets. However, the regulation is silent about the size and attribution of the dataset, which may contain revealing information. For example, if a small rural hospital releases a HIPAA-compliant dataset containing patient-level information, then recipients of the dataset may make inferences about the patients' residences (e.g., ZIP code) based on the nae and location of the hospital. If most of the patients reside in the same ZIP code, then even with no or a redacted ZIP code, a recipient of the data can still infer the full ZIP code for most patients. Similarly, while the regulation requires dates to be reported in years, more specific temporal information can sometimes be inferred. If a hospital releases a dataset daily about its emergency room visits from the day before, then exact visit date can be inferred even though only year is reported .

We use the HIPAA Safe Harbor standard to de-identify data that are not health records. Similar to how IRBs use clinical ethics to deal with all other scientific research ethics, researchers often assume that complying with HIPAA Safe Harbor requirements automatically ensures re-identification protection for their subjects and adds legal protection for themselves (by having chosen to adhere to HIPAA).

Earlier studies about vulnerabilities in HIPAA Safe Harbor data only reported unique identifications, and in so doing did not help develop scientific intuition about re-

identification risks more generally. A static value, such as the number of unique re-identifications, describes how well one re-identification strategy performed on one set of data. But how do we generalize the experience? Was it a fluke, or is it indicative of serious problems? A single number is not as useful as knowing the trajectory of small group re-identifications. How many people were re-identified as one of 2 possible named people, as one of 3 possible named people, and so on? As the number of small group re-identifications grows, so may the robustness of the re-identification risk grow.

In this study, we used the robustness of small group re-identifications to determine whether the re-identifications, even the unique ones, were likely to be correct. In Liberty Village and Bolinas, unique and small-group re-identifications resulted, but we discounted them because of the nature of the small-group re-identifications. For example, unique re-identifications and no other small group re-identifications resulted from Liberty Village data that had birth years in 10- and 20-year ranges and computer-assigned gender and race. Matching records to a homogeneous community foretells the existence of many small group re-identifications. Because none appeared, the Attackers did not believe the isolated unique re-identifications to be correct, and those re-identifications were confirmed not to be correct. Group re-identifications can be a useful aid in understanding re-identification risks.

On the other hand, re-identifications of the same kind of data for Bolinas yielded many small group re-identifications, as expected when matching records for a homogeneous community; however, the poor data quality foretells many noisy matches. Therefore, the Attackers did not believe those re-identifications were reliable, and again, the re-identifications were confirmed not to be correct.

Our results are not necessarily the worst case for re-identifying the data. Many fields in the HIPAA Safe Harbor–compliant data were unused in our experiment. Prior publications referenced chemical distinctions between the communities that were less obviously useful to the Attackers, as non-experts in environmental health. A re-identification expert may often lack domain-specific expertise that limits performance.

However, the opposite is true too: a re-identification expert may know much more than our Attackers. Data analytic companies are one of the top acquirers of publicly available hospital data [49]. Health, environmental, or legal data analytics companies whose data products benefit from re-identifying the de-identified dataset may be highly motivated and very knowledgeable about the de-identified data and therefore able to perform more re-identifications.

Our efforts did not use link analysis, deep learning or statistical matching algorithms, which are commonly used by data analytic companies to construct personal data profiles from disparate data sources [50]. Instead, we used manual and simple matching approaches. We acknowledge that more robust re-identifications may be possible using our same data and re-identification strategy with more sophisticated linking techniques.

Overall, there may exist other re-identification strategies and other data sources that may yield even more re-identifications than we demonstrated. Despite these limitations, we can state that the re-identification rate is at least as high as demonstrated here.

The original data that were the subject of our study was collected more than 10 years ago, and our effort required finding or constructing registers relevant to 2006. If the subject data had been more contemporary, then additional readily available sources of data could have been used for re-identification. For example, the HES data contained information about pets, appliances, and lawn care. If the HES data were more recent, we could have used contemporary marketing lists (e.g., [51]). While similar kinds of marketing lists existed in 2006, we were unable to obtain them retrospectively. We could have also used lookups on Facebook and other social media profiles for our re-identifications.

The environmental health researchers who conducted the original study consider the results to demonstrate a rate of re-identification risk that raises ethical cautions about sharing similar data. In prior work, Brown et al. reported on the challenges of doing community based participatory research that involves biomonitoring and household exposure studies. They emphasized the critical role consent agreements play in informing participants of potential harms as well as the steps taken to prevent or mitigate those harms [52]. They also describe the trust relationship between researchers and participants as imposing ethical requirements on researchers to protect the rights, well-being, and autonomy of participants because the participants alone may shoulder the harm based on decisions made by the researchers [54]. These considerations should be extended to include consideration of risks in datasets that are anticipated to be widely or publicly shared.

Beyond the ethical promise of anonymity, participants may suffer economic harm from loss of privacy. The value of real estate may be adversely impacted, and knowledge of research results may also impose a legal duty on the participant to inform government officials, landlords, tenants, and future homebuyers [54, 55]. This legal obligation may also result in financial costs. While all properties in a community may be impacted by outdoor air quality measures, measurements of indoor air or dust have the potential to pose greater costs to individual residents.

Environmental health studies often inform laws and regulations about industrial pollutants, which can cost companies billions of dollars (e.g., [56]). With so much money involved, protecting the identity and addresses of study participants is a critical shield from retaliatory action.

Of course, protecting privacy is not limited to cases of demonstrable economic harms. The protection of personal privacy has different goals and purposes, including upholding social values. Economic harms are often among the most dramatic examples of the consequences of loss of privacy, but other devastating consequences can be social, legal, political, and personal.

Our results show that the current HIPAA Safe Harbor cannot reliably anonymize data. How could data possibly be released with limited or virtually no risk of re-identification? To eliminate risk of re-identification, data must adhere to a formal property that provides a privacy guarantee. Computer scientists have introduced such models. The first formal protection model was *k*-anonymity, which guarantees that each record released will ambiguously map to at least *k* other records [20, 57]. Therefore, you cannot do better than

guessing $1/k$ that any particular record belongs to a named person or location. If the HES data were k-anonymized, there would be no small group re-identifications less than k and each k-sized group would be indistinguishable. This guarantee would hold regardless of the amount of redaction.

The newest formal protection model is differential privacy, which uses additive noise or subsampling to enforce a mathematical guarantee of ambiguity or disassociation [58, 59]. Unlike k-anonymity, where the actual records of the data are changed to satisfy the k requirement, a differentially private approach to smaller datasets will often make a statistical model of the original data and then produce an alternative dataset that has the same statistical properties as the original data but none of the original records. New records are generated from the statistical model itself, thereby breaking the one-to-one correspondence between records in the original and anonymized datasets. While they differ from each other, both models make provable privacy guarantees. In comparison, HIPAA Safe Harbor makes no scientific privacy guarantee.

Fifteen years ago when the HIPAA Privacy Rule was promulgated, hundreds of data brokers, offering ever-increasing amounts of personal information on Americans, did not exist. Property data and other public information were not readily available electronically. Our findings suggest that the time is ripe to modernize HIPAA Safe Harbor, especially in the face of today's data rich networked society, and to so in a manner that encourages and adopts technological innovation. Formal protection models offer the privacy guarantees that patients, and research participants, deserve.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Oscar Zarate, Laura Borth, and Sean Hooley for help with study planning, administration and logistics. This work was funded in part by a grant from the National Institute of Environmental Health Sciences (grant 1R01ES021726).

A.: Data Fields in the HIPAA Dataset

The HIPAA Dataset is a version of the original study data from the Northern California Household Exposure Study (HES) [1] that was redacted beyond the minimum requirements of the Safe Harbor provision of the Privacy Rule in the Health Information Portability and Accountability Act (HIPAA). Dates are reported within 10 or more year ranges as decades or decade groups, and all explicit geography, such as address, city and ZIP code, has been removed.

Survey File

The Survey File is a spreadsheet with 255 fields as columns and 50 data rows. Below is a description of each field. Date fields that were reported as decades are highlighted in orange

Field name	Field Description: Possible Values
PrivacyID	Unique ID created for scoring
movein	Year moved to this house: reported as decade group
housbuilt	Year house was built: reported as decade group
remod	House addition/remodeled/painted inside: Yes, No or NA
remodyr	Addition/remodel/painted inside: Before/Since beginning of year or blank
work1	Kind of remodeling work done
basement	Basement or crawl space: basement, crawl space, neither
basemfin	Is basement: Finished or Unfinished
newrug	Rugs/carpets in house new within last year: Yes, No or NA
newrugair	New rug/carpet in sample collection room: Yes, No or NA
newfurn	Large furniture new within last year: Yes, No or NA
newfurnair	New large furniture in sample collection room: Yes, No or NA
garage	Garage attached to house: Yes, No or NA
winopen	# windows open for at least 1 hr in past 24 hrs
appl_gas	Appliances or heat that use natural gas: Yes, No or NA
elecheat	Has electric heat: Yes, No or NA
natgas	Has natural gas heat: Yes, No or NA
oilheat	Has oil heat: Yes, No or NA
gas_waterht	Has gas water heater: Yes, No or NA
woodstove	Has wood stove: Yes, No or NA
usedwood	Used wood stove in past 24 hours: Yes, No or NA
keroheat	Has kerosene heater: Yes, No or NA
usedkero	Used kerosene heater in past 24 hours: Yes, No or NA
woodfire	Has wood-burning fireplace: Yes, No or NA
usedfire	Used wood-burning fireplace in past 24 hours: Yes, No or NA
gasfire	Has gas-burning fireplace: Yes, No or NA
usedgfire	Used gas-burning fireplace in past 24 hours: Yes, No or NA
goven	Has gas oven: Yes, No or NA
opergoven	Gas oven operating in past 24 hours: Yes, No or NA
gstove	Has gas stove: Yes, No or NA
opergstove	Gas stove operating in past 24 hours: Yes, No or NA
elecoven	Has electric oven: Yes, No or NA
operelecoven	Electric oven operating in past 24 hours: Yes, No or NA
elecrange	Has electric range stove: Yes, No or NA
operelrange	Electric range stove operating in past 24 hours: Yes, No or NA
fanvent	Has fan over stove: Yes, No or NA
opervent	Fan over stove operating in past 24 hours: Yes, No or NA
indgrill	Has indoor grill: Yes, No or NA
operindgrill	Indoor grill operating in past 24 hours: Yes, No or NA
dishw	Has dishwasher: Yes, No or NA
operdishw	Dishwasher operating in past 24 hours: Yes, No or NA

Field name	Field Description: Possible Values
atticfan	Has attic or window fans: Yes, No or NA
operatticfan	Attic or window fans operating in past 24 hours: Yes, No or NA
laundry	Has clothes washer in living area: Yes, No or NA
operlaund	Clothes washer in living area operating in past 24 hours: Yes, No or NA
fry_stove	Has fried food on stove: Yes, No or NA
vent_frystove	Vent fan was on while frying food: Yes, No or NA
broil_oven	Has broiled food in oven: Yes, No or NA
vent_broiloven	Vent fan was on while broiling food: Yes, No or NA
grill_indoor	Has grilled food indoors: Yes, No or NA
vent_grilling	Vent fan was on while grilling food: Yes, No or NA
bake_oven	Has baked food in oven: Yes, No or NA
vent_bakeoven	Vent fan was on while baking food: Yes, No or NA
toasteroven	Has operated a toaster or toaster oven: Yes, No or NA
vent_toaster	Vent fan was on while toasting food: Yes, No or NA
selfclean	Has cleaned the oven with self-clean heat: Yes, No or NA
vent_selfclean	Vent fan was on during the self-clean: Yes, No or NA
comp	Has computer printer: Yes, No or NA
opercomp	Computer printer operating in past 24 hours: Yes, No or NA
fax	Has fax machine: Yes, No or NA
operfax	Fax machine operating in past 24 hours: Yes, No or NA
photoc	Has photocopier: Yes, No or NA
operphoto	Photocopier operating in past 24 hours: Yes, No or NA
airfresh	Used solid air freshener in past 2 days: Yes, No or NA
airspray	Used spray air freshener in past 2 days: Yes, No or NA
hairspray	Used hair spray in past 2 days: Yes, No or NA
antipers	Used spray antiperspirant in past 2 days: Yes, No or NA
deterg	Used laundry detergent in past 2 days: Yes, No or NA
dishdeterg	Used dishwasher detergent in past 2 days: Yes, No or NA
surfclean	Used spray-on surface cleaner in past 2 days: Yes, No or NA
ovenclean	Used oven cleaner in past 2 days: Yes, No or NA
glues	Used glues or adhesives in past 2 days: Yes, No or NA
furnpol	Used furniture polish in past 2 days: Yes, No or NA
toilclean	Used toilet cleaner in past 2 days: Yes, No or NA
tileclean	Used tub or tile cleaner in past 2 days: Yes, No or NA
painthin	Used paint thinner/stripper in past 2 days: Yes, No or NA
bugkill	Used bug killers/pesticides in past 2 days: Yes, No or NA
carpclean	Used carpet cleaner in past 2 days: Yes, No or NA
spotrem	Used spot remover in past 2 days: Yes, No or NA
mothball	Used mothballs in past 2 days: Yes, No or NA
nailpol	Used fingernail polish in past 2 days: Yes, No or NA
homebus	Business operating in house: Yes, No or NA
homehobb	Workshop/hobby area in house: Yes, No or NA

Field name	Field Description: Possible Values
everspray	House ever been treated for bugs: Yes, No or NA
sprayear	House treated for bugs in past year: Reported as decade
recentmth_treat	Most recent month treated for bugs during past year: Jan,…,Dec
numb_mthsrx	Number of months of treatment reported
sprayreason	Kind of bugs house treated for in past year
spraytype	What was house treated for bugs with in past year
treat_cat	Treatment category: Spray/Exterminator/Bomb, Bait/Stake/Borax, NoInfo
lastspray	Most recent year house treated for bugs: Reported as decade
lastsprayreason	Kind of bugs house treated for in most recent year
lastspraytype	What was house treated for bugs with in most recent year
lastsprayamt	How many times house treated for bugs in most recent year
sprayother	House treated for bugs any other years: None or NA
sprayothyr1a	Bug treatment 1, year 1: Reported as decade
sprayothyr1b	Bug treatment 1, year 2: Reported as decade
sprayothbug1	Bug treatment 1, kind of bugs treated for
sprayothtype1	Bug treatment 1, treated with
sprayothamt1	Bug treatment 1, how often
sprayothyr2a	Bug treatment 2, year 1: Reported as decade
sprayothyr2b	Bug treatment 2, year 2: Reported as decade
sprayothbug2	Bug treatment 2, kind of bugs treated for
sprayothtype2	Bug treatment 2, treated with
sprayothamt2	Bug treatment 2, how often
sprayothyr3a	Bug treatment 3, year 1: Reported as decade
sprayothyr3b	Bug treatment 3, year 2: Reported as decade
sprayothbug3	Bug treatment 3, kind of bugs treated for
sprayothtype3	Bug treatment 3, treated with
sprayothamt3	Bug treatment 3, how often
sprayothyr4a	Bug treatment 4, year 1: Reported as decade
sprayothyr4b	Bug treatment 4, year 2: Reported as decade
sprayothbug4	Bug treatment 4, kind of bugs treated for
sprayothtype4	Bug treatment 4, treated with
sprayothamt4	Bug treatment 4, how often
sprayothyr5a	Bug treatment 5, year 1: Reported as decade
sprayothyr5b	Bug treatment 5, year 2: Reported as decade
sprayothbug5	Bug treatment 5, kind of bugs treated for
sprayothtype5	Bug treatment 5, treated with
sprayothamt5	Bug treatment 5, how often
treatment_cat	Treatment category: Spray/Exterminator/Bomb, Bait/Stake/Borax, NoInfo
lawncare	Who cares for lawn: Household or Building management
lawntx	Lawn ever treated with insecticide or herbicide: Yes, No, Don’t know
lawnyr	Lawn treated in past year: Yes, No, Don’t know
lawnmonth	Most recent month lawn treated for bugs in last year: Jan,…,Dec

Field name	Field Description: Possible Values
lawnmonth_num	Number of months lawn treated in last year
lawnreas	What was lawn treated for in past year
lawntype	What was lawn treated with in past year
lawntreat_cat	Lawn treatment category: Spray/Exterminator/Bomb, Bait/Stake/Borax
lawnyrlast	Most recent year lawn was treated: Reported as decade
lawnlastreas	What was lawn treated for in most recent year
lawnlastype	What was lawn treated with in most recent year
lawnoth	Lawn treated any other years: None or NA
lawnothyr1a	Lawn treatment 1, year 1: Reported as decade
lawnothyr1b	Lawn treatment 1, year 2: Reported as decade
lawnothreas1	Lawn treatment 1, treated for
lawnothtype1	Lawn treatment 1, treated with
lawnothyr2a	Lawn treatment 2, year 1: Reported as decade
lawnothyr2b	Lawn treatment 2, year 2: Reported as decade
lawnothreas2	Lawn treatment 2, treated for
lawnothtype2	Lawn treatment 2, treated with
lawnothyr3a	Lawn treatment 3, year 1: Reported as decade
lawnothyr3b	Lawn treatment 3, year 2: Reported as decade
lawnothreas3	Lawn treatment 3, treated for
lawnothtype3	Lawn treatment 3, treated with
lawnothyr4a	Lawn treatment 4, year 1: Reported as decade
lawnothyr4b	Lawn treatment 4, year 2: Reported as decade
lawnothreas4	Lawn treatment 4, treated for
lawnothtype4	Lawn treatment 4, treated with
lawntreat_cat2	Lawn treatment category for other years: Spray/…/Bomb, Bait/…/Borax
lawnpro	Ever used professional lawn care service: Yes, No, Don’t know
lawnprolast	Most recent year used professional lawn service: Since start or NA
lawnprofirst	First year used professional lawn service:Reported as decade
lawnprooth	Used professional lawn service any other years: Yes, No, Don’t know
lawnproyr1a	Professional lawn service 1, year 1: Reported as decade
lawnproyr1b	Professional lawn service 1, year 2: Reported as decade
lawnproyr2a	Professional lawn service 2, year 1: Reported as decade
lawnproyr2b	Professional lawn service 2, year 2: Reported as decade
lawnproyr3a	Professional lawn service 3, year 1: Reported as decade
lawnproyr3b	Professional lawn service 3, year 2: Reported as decade
lawnproyr4a	Professional lawn service 4, year 1: Reported as decade
lawnproyr4b	Professional lawn service 4, year 2: Reported as decade
pets	Any pets: Yes or No
fleatx	Any cats and/or dogs treated for fleas: Yes or No
fleatype	What type of flea treatment
flealast_mon	Most recent flea treatment – month: Jan,…,Dec
totalsquareft	Total square feet of living area in ranges: 450-500, 650-700, 700-1000, …

Field name	Field Description: Possible Values
smoke	Anyone who lives in house smoke tobacco: Yes or No
smoke24	Anyone who lives in house smoke tobacco in past 24: Yes or No
room1samp	Room #1 - sampled: Checked or Not checked
room1	Room #1 - type: Living room, Kitchen, Bedroom, Bath, Other
norug1	Room #1 - no rug: Checked or Not checked
vinylfl1	Room #1 - vinyl flooring: Yes or No
rug1age	Room #1 - age of rug/carpet (yrs): Reported as decade
rug1wall	Room #1 - wall to wall carpet: Yes or No
rug1half	Room #1 - area rug > 1/2 of room: Yes or No
room2samp	Room #2 - sampled: Checked or Not checked
room2	Room #2 - type: Living room, Kitchen, Bedroom, Bath, Other
norug2	Room #2 - no rug: Checked or Not checked
vinylfl2	Room #2 - vinyl flooring: Yes or No
rug2age	Room #2 - age of rug/carpet (yrs) : Reported as decade
rug2wall	Room #2 - wall to wall carpet: Yes or No
rug2half	Room #2 - area rug > 1/2 of room: Yes or No
room3samp	Room #3 - sampled: Checked or Not checked
room3	Room #3 - type: Living room, Kitchen, Bedroom, Bath, Other
norug3	Room #3 - no rug: Checked or Not checked
vinylfl3	Room #3 - vinyl flooring: Yes or No
rug3age	Room #3 - age of rug/carpet (yrs) : Reported as decade
rug3wall	Room #3 - wall to wall carpet: Yes or No
rug3half	Room #3 - area rug > 1/2 of room: Yes or No
room4samp	Room #4 - sampled: Checked or Not checked
room4	Room #4 - type: Living room, Kitchen, Bedroom, Bath, Other
norug4	Room #4 - no rug: Checked or Not checked
vinylfl4	Room #4 - vinyl flooring: Yes or No
rug4age	Room #4 - age of rug/carpet (yrs) : Reported as decade
rug4wall	Room #4 - wall to wall carpet: Yes or No
rug4half	Room #4 - area rug > 1/2 of room: Yes or No
room5samp	Room #5 - sampled: Checked or Not checked
room5	Room #5 - type: Living room, Kitchen, Bedroom, Bath, Other
norug5	Room #5 - no rug: Checked or Not checked
vinylfl5	Room #5 - vinyl flooring: Yes or No
rug5age	Room #5 - age of rug/carpet (yrs) : Reported as decade
rug5wall	Room #5 - wall to wall carpet: Yes or No
rug5half	Room #5 - area rug > 1/2 of room: Yes or No
room6samp	Room #6 - sampled: Checked or Not checked
room6	Room #6 - type: Living room, Kitchen, Bedroom, Bath, Other
norug6	Room #6 - no rug: Checked or Not checked
vinylfl6	Room #6 - vinyl flooring: Yes or No
rug6age	Room #6 - age of rug/carpet (yrs) : Reported as decade

Field name	Field Description: Possible Values
rug6wall	Room #6 - wall to wall carpet: Yes or No
rug6half	Room #6 - area rug > 1/2 of room: Yes or No
room7samp	Room #7 - sampled: Checked or Not checked
room7	Room #7 - type: Living room, Kitchen, Bedroom, Bath, Other
norug7	Room #7 - no rug: Checked or Not checked
vinylfl7	Room #7 - vinyl flooring: Yes or No
rug7age	Room #7 - age of rug/carpet (yrs) : Reported as decade
rug7wall	Room #7 - wall to wall carpet: Yes or No
rug7half	Room #7 - area rug > 1/2 of room: Yes or No
room8samp	Room #8 - sampled: Checked or Not checked
room8	Room #8 - type: Living room, Kitchen, Bedroom, Bath, Other
norug8	Room #8 - no rug: Checked or Not checked
vinylfl8	Room #8 - vinyl flooring: Yes or No
rug8age	Room #8 - age of rug/carpet (yrs) : Reported as decade
rug8wall	Room #8 - wall to wall carpet: Yes or No
rug8half	Room #8 - area rug > 1/2 of room: Yes or No
room9samp	Room #9 - sampled: Checked or Not checked
room9	Room #9 - type: Living room, Kitchen, Bedroom, Bath, Other
norug9	Room #9 - no rug: Checked or Not checked
vinylfl9	Room #9 - vinyl flooring: Yes or No
rug9age	Room #9 - age of rug/carpet (yrs) : Reported as decade
rug9wall	Room #9 - wall to wall carpet: Yes or No
rug9half	Room #9 - area rug > 1/2 of room: Yes or No
room10samp	Room #10 - sampled: Checked or Not checked
room10	Room #10 - type: Living room, Kitchen, Bedroom, Bath, Other
norug10	Room #10 - no rug: Checked or Not checked
vinylfl10	Room #10 - vinyl flooring: Yes or No
rug10age	Room #10 - age of rug/carpet (yrs) : Reported as decade
rug10wall	Room #10 - wall to wall carpet: Yes or No
rug10half	Room #10 - area rug > 1/2 of room: Yes or No
birth_yr	Year born in: Reported as decade
job	Working at a job: Yes or No
jobtype	Type of job: by industry
other_job	Anyone else in the house working at a job: Yes or No
other_jobtype	What kind of job is the other person working at: by industry
school	Highest grade in school completed: <= 8th grade, Some high school,…
own_home	Owns own home: Yes or No
race_white	White: White or NA
race_black	Black: Black or NA
race_his	Hispanic: Hispanic or NA
race_nam	Native American: Native American or NA
race_asian	Asian: Asian or NA

Field name	Field Description: Possible Values
race_other	Other: Something else or NA
race_otherspe	Other race-specified
sex	Gender of respondent: Male or Female
SurveyLanguage	Survey language: English or Spanish

Air and Dust Files

Field	Description
Compound	Compound name; naming conventions use Chemlist file
Concentration	specific MRLs for dust and air.
Flag	Data flag. 1 = Detect; 0 = Non-detect; 0.5 and 0.6 = estimated value
Units	reporting units
Privacy.ID	unique participant identifier; re-coded from original values
Media	sampling media
Analyte.MRL	Compound-specific method reporting limit

B.: HES Published Toxicology

Below is a reprint of selected summary statistics for the outdoor air results from the Household Exposure Study published as supporting information for an academic paper about the investigation [17]. The full table is available at http://pubs.acs.org/doi/full/10.1021/es100159c. The attackers used outdoor fluoranthene levels to distinguish homes in Richmond from homes in Bolinas.

Biographies

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, Editor-in-Chief of Technology Science, and was formerly the Chief Technology Officer of the U.S. Federal Trade Commission. She earned her PhD in computer science from the Massachusetts Institute of Technology and her undergraduate degree from Harvard. More information about Dr. Sweeney is available at her website at latanyasweeney.org. As Editor-In-Chief of *Technology Science*, Professor Sweeney was recused from the review of this paper.

Ji Su Yoo is an experienced researcher at Harvard University's Data Privacy Lab and at the Institute for Quantitative Social Science. She graduated from Harvard College with a B.A. in Social Studies. Currently, she is interested in the intersection between government and technology and how their interaction can impact privacy and data analysis in health care. She is also interested in exposing how technology and its use in government and policy-making may exacerbate or reflect existing political and societal inequalities. Ji Su hopes to attend graduate school and 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.

Laura Perovich is a PhD student at the Media Lab at the Massachusetts Institute of Technology (MIT). Her work focuses on environmental health, human computer interaction, and data visualization. She was previously a researcher at Silent Spring Institute and was involved in considering how to share Household Exposure Study data with the U.S. Environmental Protection Agency. She holds a S.M. from MIT and a B.A. from Bowdoin College.

Katie Boronow is a staff scientist at Silent Spring Institute, a scientific research organization that studies environmental chemicals and women's health, with a particular focus on breast cancer. Her research leverages digital tools for reporting personal exposure results to participants in biomonitoring studies as a platform for increasing knowledge about endocrine disrupting chemicals. She holds a master's degree from Harvard University in organismic and evolutionary biology and a bachelor's degree summa cum laude from Yale University in biology.

Phil Brown is University Distinguished Professor of Sociology and Health Science at Northeastern University, where he directs the Social Science Environmental Health Research Institute www.northeastern.edu/environmentalhealth He is the author of No Safe Place: Toxic Waste, Leukemia, and Community Action, and Toxic Exposures: Contested Illnesses and the Environmental Health Movement, and co-editor of Social Movements in Health, and Contested Illnesses: Citizens, Science and Health Social Movements. He studies biomonitoring and household exposure and reporting back data to participants in collaboration with Silent Spring Institute (silentspring.org), social policy concerning flame retardants and perfluorinated compounds (pfasproject.com), and health social movements. He directs an NIEHS T-32 training program, "Transdisciplinary Training at the Intersection of Environmental Health and Social Science." He heads the Community Outreach and Translation Core of Northeastern's Children's Environmental Health Center (Center for Research on Early Childhood Exposure and Development in Puerto Rico/CRECE) www.northeastern.edu/crece and both the Research Translation Core and Community Engagement Core of Northeastern's Superfund Research Program (Puerto Rico Testsite to Explore Contamination Threats (PROTECT) www.northeastern.edu/protect. He is on the National Advisory Environmental Health Science Council, which advises the director of NIH's National Institute of Environmental Health Sciences.

Julia Green Brody, Ph.D., is executive director and senior scientist at Silent Spring Institute, a scientific research organization that studies environmental chemicals and women's health, with a particular focus on breast cancer. She co-led the Household Exposure Study, which was described in Environmental Science & Technology as the most comprehensive study of endocrine disrupting compounds in homes. The study was the first to show that consumer products and indoor environments are major sources of exposure to phthalates, phenols, and flame retardants, among other chemicals. The value of this dataset led to her interest in the potential privacy risks of sharing environmental exposure data and to her partnership with Dr. Sweeney. Dr. Brody's research has been funded by the National Institutes of Health and

National Science Foundation and was recognized by the U.S. Environmental Protection Agency with an Environmental Merit Award in 2000. Dr. Brody is an adjunct assistant professor at the Brown University School of Medicine. She earned her Ph.D. at the University of Texas at Austin.

References

1. Brody J, Morello-Frosch R, Zota A, Brown P, Pérez C, Rudel R. Linking Exposure Assessment Science With Policy Objectives for Environmental Justice and Breast Cancer Advocacy: The Northern California Household Exposure Study. American Journal of Public Health. 2009;99(Suppl 3):S600–S609. doi:10.2105/AmericanJournalofPublicHealth.2008.149088. http://www-ncbi-nlm-nih-gov.ezp-prod1.hul.harvard.edu/pmc/articles/PMC2774181 [PubMed: 19890164]
2. Kwok P and Lafky D. Harder Than You Think: A case study of re-identification risk of HIPAA-compliant records. https://www.researchgate.net/publication/265077763_Harder_Than_You_Think_A_Case_Study_of_Re-identification_Risk_of_HIPAA-Compliant_Records
3. Sweeney L Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3 Pittsburgh 2000 http://dataprivacylab.org/projects/identifiability/
4. U.S. Health Insurance Portability and Accountability Act of 1996. 45 CFR Parts 160 and 164. 2 2003
5. U.S. Department of Health and Human Services. Summary of the HIPAA Privacy Rule. Accessed March 10, 2017. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/
6. U.S. Health Insurance Portability and Accountability Act of 1996. Safe Harbor. 45 CFR 164(b)(2)(ii).
7. U.S. Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. 11 26, 2012 Accessed March 10, 2017 https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/#safeharborguidance
8. U.S. Department of Health and Human Services. NPRM for Revisions to the Common Rule. Federal Register. 80(173) pp. 53933–54061. 9 8, 2015 https://www.gpo.gov/fdsys/pkg/FR-2015-09-08/pdf/2015-21756.pdf
9. Sweeney L Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903 9 29, 2015 http://techscience.org/a/2015092903
10. Federal Committee on Statistical Methodology. Report on Statistical Disclosure Limitation Methodology. Statistical Working Paper 22. 12 2005 http://www.hhs.gov/sites/default/files/spwp22.pdf]
11. El Emam K, Jonker E, Arbuckle L, and Malin B. “A Systematic Review of Re-Identification Attacks on Health Data.” PLoS ONEOne, vol. 6, no. 12, 12 2011, pp. 1–12.
12. Narayanan A and Shmatikov V. “Robust De-anonymization of Large Spare Datasets. Proceedings of the IEEE Symposium on Security and Privacy, 2008, pp. 111–125
13. PLOS PLoS One. Data Availability. http://journals.plos.org/plosone/s/data-availability
14. Journal Open Data Policies. http://oad.simmons.edu/oadwiki/Journal_open-data_policies
15. U.S. Federal Government. Open Government. Accessed March 2015 https://www.data.gov/open-gov/
16. Adams C, Brown P, Morello-Frosch R, et al. Disentangling the Exposure Experience: The Roles of Community Context and Report-back of Environmental Exposure Data. Journal of health and social behavior. 2011;52(2):180–196. doi:10.1177/0022146510395593. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3175404/ [PubMed: 21673146]
17. Rudel R, Dodson R, Perovich L, Morello-Frosch R, et al. Semivolatile Endocrine-Disrupting Compounds in Paired Indoor and Outdoor Air in Two Northern California Communities. Environmental Science and Technology 44 (17) 2010 http://pubs.acs.org/doi/full/10.1021/es100159c

18. Zota A, Rudel R, Morello-Frosch R and Brody J. Elevated House Dust and Serum Concentrations of PBDEs in California: unintended consequences of furniture flammability standards? Environmental Science and Technology 42 (21) 2008.
19. Dunagan S, Dodson R, Rudel R and Brody J. Toxics Use Reduction in the Home: lessons learned from household exposure studies. Journal of Cleaner Production March 2011 19(5): 438–444. doi: 10.1016/j.jclepro.2010.06.012
20. Sweeney L k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557–570. http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.html
21. Sweeney L Patient Identifiability in Pharmaceutical Marketing Data. Data Privacy Lab Working Paper 1015. Cambridge 2011 https://dataprivacylab.org/projects/identifiability/pharma1.html
22. Alexander L and Jabine T Access to social security microdata files for research and statistical purposes. Social Security Bulletin. 1978 (41) No. 8.
23. California Department of Health Services. Public Aggregate Reporting – Guidelines Development Project. Version 1.6 8 25, 2014 http://www.dhcs.ca.gov/dataandstats/data/DocumentsOLD/IMD/PublicReportingGuidelines.pdf
24. Romanos A Defamation Update (New Zealand Defamation Law). 2014 http://www.defamationupdate.co.nz/guide-to-defamation-law
25. Who can Sue for Defamation Digital Media Law Project. Berkman Center for Internet and Society Harvard University http://www.dmlp.org/legal-guide/who-can-sue-defamation
26. Ross S Deciding Communication Law: Key Cases in Context. Lawrence Erlbaun Associates Mahwah, New Jersey 2004 p507 https://www.amazon.com/Deciding-Communication-Law-Context-Routledge/dp/0415647150
27. Sweeney L Privacert Risk Assessment Server. 2004 http://privacert.com/
28. Dodson R, Perovich L, Covaci A, Van den Eede N, Ionas A, Dirtu A, Brody J, Rudel R. After the PBDE phase-out: A broad suite of flame retardants in repeat house dust samples from California. Environmental Science & Technology, 2012; 46(24):13056–13066. doi:10.1021/es303879n [PubMed: 23185960]
29. County of Contra Costa Assessor's Office. Copies of Records Containing Tax Information for 2006. County of Contra Costa, California 12 2013 http://www.co.contra-costa.ca.us/191/Assessor
30. County of Marin Assessor's Office. Copies of Records Containing Tax Information for 2006. County of Marin, California 12 2015 https://www.marincounty.org/depts/ar/divisions/assessor
31. World Privacy Forum. Data Brokers Opt Out List. 1 2017 https://www.worldprivacyforum.org/2013/12/data-brokers-opt-out/
32. Angwin J Privacy Tools: Opting out from data brokers. Propublica. 1 30, 2014 https://www.propublica.org/article/privacy-tools-opting-out-from-data-brokers
33. U.S. Federal Trade Commission. A Call for Transparency and Accountability. 5 2014 https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf
34. Google Earth. https://www.google.com/earth/
35. U.S. Bureau of the Census. First names by Gender from the 1990 Census. Accessed March 2017 https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html
36. U.S. Census Bureau. Frequently Occurring Surnames from the Census 2000. 9 15, 2014 https://www.census.gov/topics/population/genealogy/data/2000_surnames.html
37. Python Programming Language. https://www.python.org/
38. Marin County Tax Assessor. 2006 Equalized Secured Roll Bolinas Area. County of Marin, California http://www.marincounty.org/residents/your-home/taxes-and-assessments
39. Marion County Tax Assessor. Search Assessor Records by Parcel Number. County of Marin, California https://www.marincounty.org/depts/ar/divisions/assessor/search-assessor-records
40. Colliers International. Liberty Village: For Sale Investment Summary. Accessed May 2016 http://www.colliers.com/~/media/Files/United%20States/MARKETS/San%20Francisco/Featured%20Properties/Liberty%20Village%20Flyer

41. Liberty Village. LoopNet. Accessed May 2016 http://www.loopnet.com/Listing/18369851/298-West-Chanslor-Avenue-Richmond-CA/
42. Marion County Tax Assessor. Search Assessor Records by Parcel Number. County of Marin, California http://apps.marincounty.org/TaxRollSearch1/
43. Page and Turnbull. Atchison Village: Mini-historic Structure Report and Preservation Plan. 9 30, 2009 http://www.ci.richmond.ca.us/DocumentCenter/Home/View/5791
44. U.S. Department of Health and Human Services. Standards for Privacy of Individually Identifiable Health Information; Proposed Rule. 45 CFR Parts 160 through 164. Federal Register. 64(212) 11 3, 1999.
45. Dirmark Media. Ailments Mailing Lists. Accessed March 2017 http://dmdatabases.com/databases/consumer-mailing-lists/ailments-lists
46. Sweeney, ed. The DataMap: Documenting all the places personal data goes. Accessed March 2017 http://thedatamap.org
47. Sweeney L Only You, Your Doctor, and Many Others May Know. Technology Science. 2015092903 9 29, 2015 https://techscience.org/a/2015092903
48. Zang J, Dummit K, Graves J, Lisker P, Sweeney L. Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps. Technology Science. 2015103001 10 30, 2015 https://techscience.org/a/2015103001
49. Robertson J Public Records Requests for State Discharge Data (updated with Maine). Bloomberg News. ForeverData.org. Collection 1007. 11 2012 https://foreverdata.org/1007/
50. Data Analytics. Technopedia. 5 2017 https://www.techopedia.com/definition/26418/data-analytics
51. Dirmark Media. Consumer Mailing Lists. Accessed March 2017 http://dmdatabases.com/databases/consumer-mailing-lists
52. Brown P, Morello-Frosch R, Brody J, et al. Institutional Review Board Challenges Related to Community-Based Participatory Research on Human Exposure to Environmental Toxins: a case study. Environmental Health. 2010 9(39). http://www.ehjournal.net/content/9/1/39
53. Cordner A, Ciplet D, Brown P and Morello-Frosch R. Reflexive Research Ethics for Environmental Health and Justics: academics and movement-building. Social Movement Studies. 2012; 11(2): 161–176. doi:10.1080/14742837.2012.664898. [PubMed: 22690133]
54. Goho S 2016 The legal implications of report back in household exposure studies. Environmental Health Perspectives 124:1662–1670; 10.1289/EHP187 https://ehp.niehs.nih.gov/ehp187/ [PubMed: 27153111]
55. Resnik DB. 2012 Environmental Health Ethics. Cambridge, UK: Cambridge University Press.
56. Chow L $90 Billion Whistleblower Suit Filed Against Four of the Nation's Largest Chemical Companies. EcoWatch. 9 16, 2016 http://www.ecowatch.com/whistleblower-lawsuit-chemical-companies-2005784783.html
57. 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
58. Dwork C. Differential Privacy: a survey of results; International Conference on Theory and Applications of Models of Computation TAMC; 2008. 1–19.
59. Dwork C The Differential Privacy Frontier. International Association for Cryptologic Research. TCC 2009.

Highlights

- The HIPAA Safe Harbor is not sufficient to protect data against re-identification
- We found correct re-identifications for ~25% of records in a subset of a HIPAA-compliant environmental health dataset
- We used demographic and non-demographic fields to link a HIPAA-compliant dataset with external data sources
- Group re-identifications can extend potential harms to all individuals associated with the same record

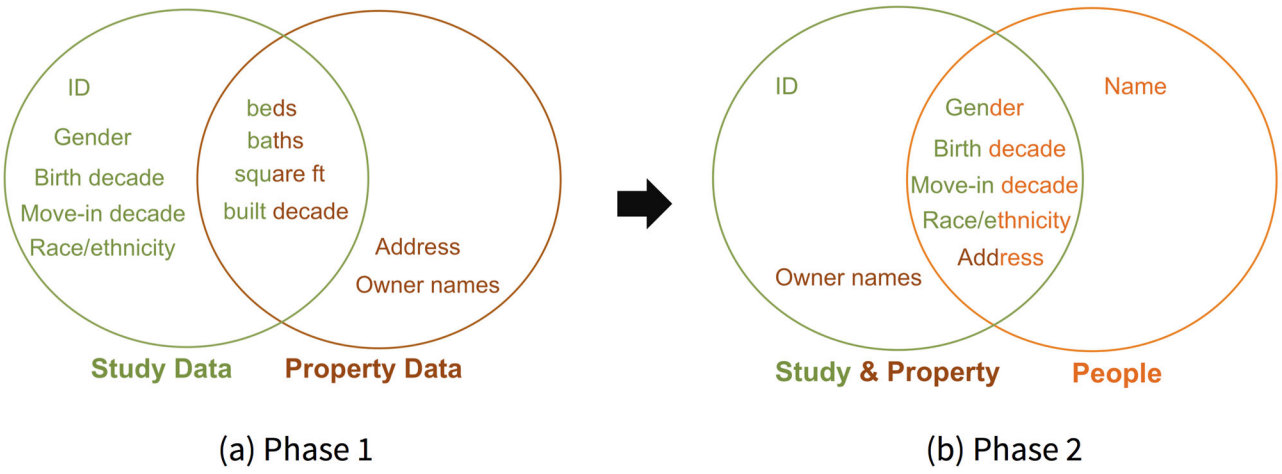


Figure 1.

Re-identification strategy to associate an ID in the Study Data with an Address and Name of a participant in the study. (a) First, match records in the HES study data (green) to real estate property data (brown) on beds, baths, square feet of living area, and decade group in which the house was built (mixed green and brown lettering) to put an address to an ID. Then, (b) match the joined information (green and brown) from (a) on gender, decade group of birth, move-in decade group, race and ethnicity, and address (mixed green, orange, and brown lettering) to a people register (orange) to put a name to the ID. The name of a known resident at the address in the people register may or may not be the same as the name of the owners in the real estate property data. A re-identification results when a name or address is associated with an ID. “Decade” refers to dates grouped in ranges of 10 to 90 years.

Field name	Field Description: Possible Values
PrivacyID	From HIPAA Dataset: Unique ID created for scoring
birth_yr	From HIPAA Dataset: Year born in: reported as decade group
school	From HIPAA Dataset: Highest grade in school completed: <= 8th grade, …
own_home	From HIPAA Dataset: Owns own home: Yes or No
gender	From HIPAA Dataset: Gender of respondent: Male or Female
movein	From HIPAA Dataset: Year moved to this house: as decade group
housebuilt	From HIPAA Dataset: Year house was built: reported as decade group
SurveyLanguage	From HIPAA Dataset: Survey language: English or Spanish
totalsquareft	From HIPAA Dataset: Total square feet of living area in ranges
race_asian	From HIPAA Dataset: Asian or NA
race_white	From HIPAA Dataset: White or NA
race_black	From HIPAA Dataset: Black or NA
race_his	From HIPAA Dataset: Hispanic or NA
race_nam	From HIPAA Dataset: Native American or NA
fluoranthene	Copied from air data (Fluoranthene) in HIPAA Dataset
totalrooms	Computed from room1,…,room9 fields in HIPAA Dataset
numberBaths	Computed from room1,…,room9 fields in HIPAA Dataset
numberBeds	Computed from room1,…,room9 fields in HIPAA Dataset

Figure 2.

Fields of the De-ID Dataset, as selected and computed from fields in the HIPAA Dataset listed in Appendix A. These 18 fields are the basis for re-identification. Possible values for birth_yr: 1920-1939, 1940-1949, 1950-69, or 1970-1989; for move-in: 1970-1989, 1990-1999, or 2000-2009; housebuilt: 1840-1949 or 1950-1989; and for totalsquareft: 450-500, 500-650, 650-700, 700-1000, or 1000-2000.

Field name	Field Description (original field names in tax assessor data)
owner1	Name of primary owner of property (P_OWNR_NM)
owner2	Name of secondary owner of property (OWN_NM_2ND)
address	Address of property (concatenate S_STR_NBR, S_STR_NM, S_STR_SUF)
numberBeds	Number of bedrooms (BEDS)
numberBaths	Number of bathrooms (BATHS)
totalrooms	Total number of rooms (TOT_ROOMS)
housebuilt	Year house built (YR_HS_BLT)
totalsquareft	Total living area (TLA)

Beds	Baths	Total Rooms	Living Area	Number of Units	Percent
1	1	3	554	68	15%
1	1	3	851	1	0%
2	1	4	672	47	10%
2	1	4	781	179	40%
2	1	4	787	36	8%
2	1	4	799	1	0%
2	1	4	799	2	0%
2	1	4	851	2	0%
2	1	4	865	1	0%
2	1	4	922	2	0%
2	1	5	1003	1	0%
2	1	5	1081	1	0%
2	1	5	736	1	0%
2	1	5	792	1	0%
2	1	5	851	2	0%
2	1.5	4	907	1	0%
2	1.5	4	925	1	0%
2	2	4	937	1	0%
3	1	5	851	47	10%
3	1	5	865	48	11%
3	1	6	1007	1	0%
3	1	6	885	1	0%
3	1	6	977	1	0%
3	1	7	1234	1	0%
3	1.5	6	971	1	0%
3	2	6	1188	1	0%
4	1	5	985	1	0%

Figure 3a.

Fields of the Atchison Village Property Register, which are a subset of fields selected from the 2006 tax assessor data for Atchison Village, as acquired from the County of Contra Costa Assessor's Office [29].

Figure 3b. The number and percent of units in Atchison Village having specific combinations of bedrooms, bathrooms, total rooms, and total living area. There are a total of 450 units. The typical unit has 2 bedrooms and 1 bath (179 of 450 units or 40 percent).

Figure 4.
Original Google aerial image of Liberty Village. The brown-roofed buildings comprise Liberty Village [34].

Figure 5.

Measurements of the rooflines of the buildings in Liberty Village. Three measurements found: red was 1.9 cm, green was 2.1 cm, and blue was 2.6 cm. Red lines show duplexes with 2 one-bedroom units. Green lines show duplexes with 2 two-bedroom units. Blue lines show duplexes with 2 three-bedroom units. In total, there are 100 units.

Figure 6.

Addresses of the 100 units in Liberty Village (rectangles) and the parking spot numbers (yellow circles) as inferred by the Attackers. Streets are Chanslor Circle (CC), Chanslor Row (Row), Circle Court (Circle Ct), and West Chanslor Avenue (W).

Beds	Baths	Total Rooms	Living Area	Number	Percent
1	1	4	528	8	7%
2	1	5	624	75	67%
3	1	6	816	29	26%

Figure 7.

The number and percentage of units in Liberty Village having specific combinations of bedrooms, bathrooms, total rooms, and total living area. A total of 100 units were found with 111 possible addresses. The typical unit has 2 bedrooms and 1 bath (75 of 111 units or 68 percent).

Field name	Field Description: and possible values
address	Address on: Circle Ct, W Chanslor Ave, or Chanslor Cir, Row, or Ave
numberBeds	Number of bedrooms: 1, 2, or 3
numberBaths	Number of bathrooms: 1
totalrooms	Total number of rooms: 4, 5, or 6
housebuilt	Year house built: 1942
totalsquareft	Total living area: 528 sq ft, 624 sq ft, or 816 sq ft

Figure 8.
Fields of the Liberty Village Property Register constructed by the Attackers using aerial images and published facts about the rental community.

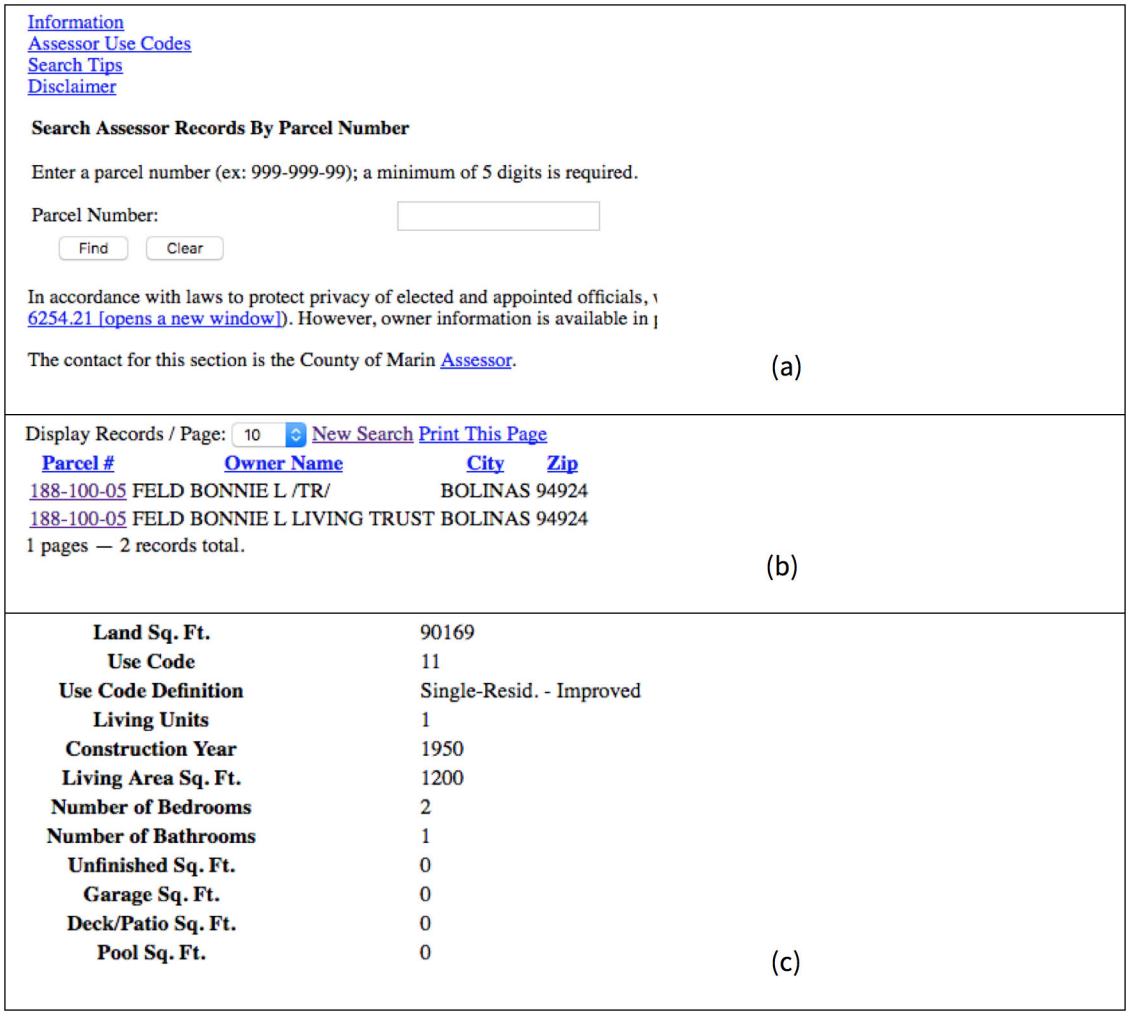


Figure 9.

A walk-through of the Marin County Tax Assessor website [42] to learn the housing characteristics for a Bolinas property: (a) The initial screen requires a parcel number, which appears as the property id in the Bolinas Tax Data. (b) Search results for the property id (or parcel) 188-100-05. (c) Selection of a search result gives housing characteristics for the property id, in this case a single residential home constructed in 1950 having 2 bedrooms and 1 bath in a living area of 1,200 square feet.

Bedrooms	Total	Percent
0	3	1%
1	93	18%
2	232	44%
3	131	25%
4	46	9%
5	15	3%
6	5	1%
7	3	1%
8	1	0%
9	1	0%
Baths	**Total**	**Percent**
0	2	0%
1	281	53%
1.5	33	6%
2	137	26%
2.5	27	5%
3	28	5%
3.5	8	2%
4	9	2%
5	3	1%
5.5	2	0%
Total Living Area	**Square ft.**	
Smallest	274	
Largest	8,093	
Median	1,311	
Average	1,496	
Standard Deviation	815	
Year House Built		
Earliest	1879	
Latest	2005	
Median	1959	
Average	1929	
Standard Deviation	208	
Value not known	6	
Total Units	**530**	

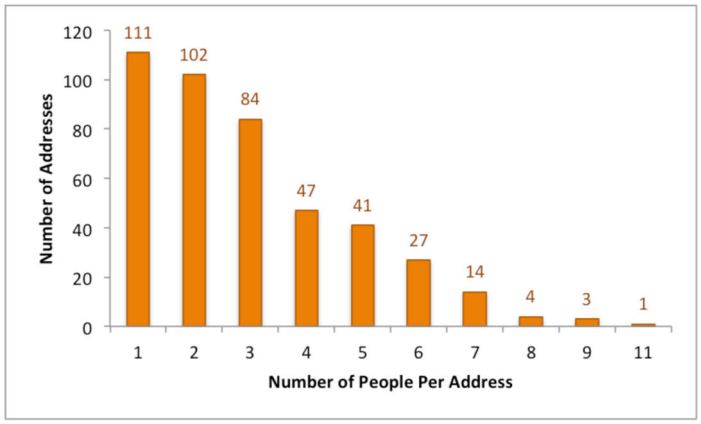
Figure 10.

Number and percentage of bedrooms and baths, and statistics about the living space and year houses were built in Bolinas, based on a total of 530 addresses having descriptive tax assessor data [38, 39, 42].

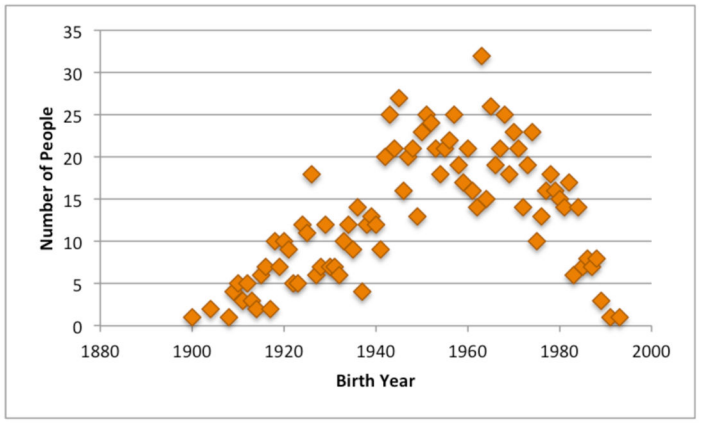
Field name	Field Description (original field names in tax assessor data)
owner1	Name of primary owner of property
owner2	Name of secondary owner of property
address	Address of property
numberBeds	Number of bedrooms
numberBaths	Number of bathrooms
housebuilt	Year house built
totalsquareft	Total living area
garagesqft	Total square feet of garage

Figure 11.

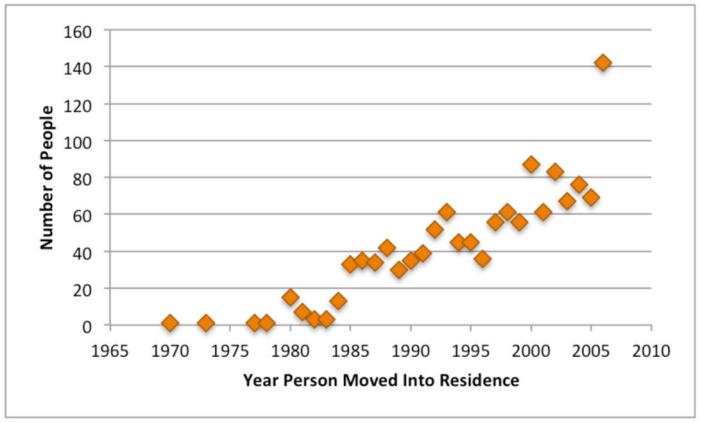
Fields of the Bolinas Property Register. Addresses and ownership based on 2006 tax assessor data, as acquired from the Marin County Assessor's Office [38]. Housing characteristics from the Marin County Assessor's Website in 2013 [39, 42].



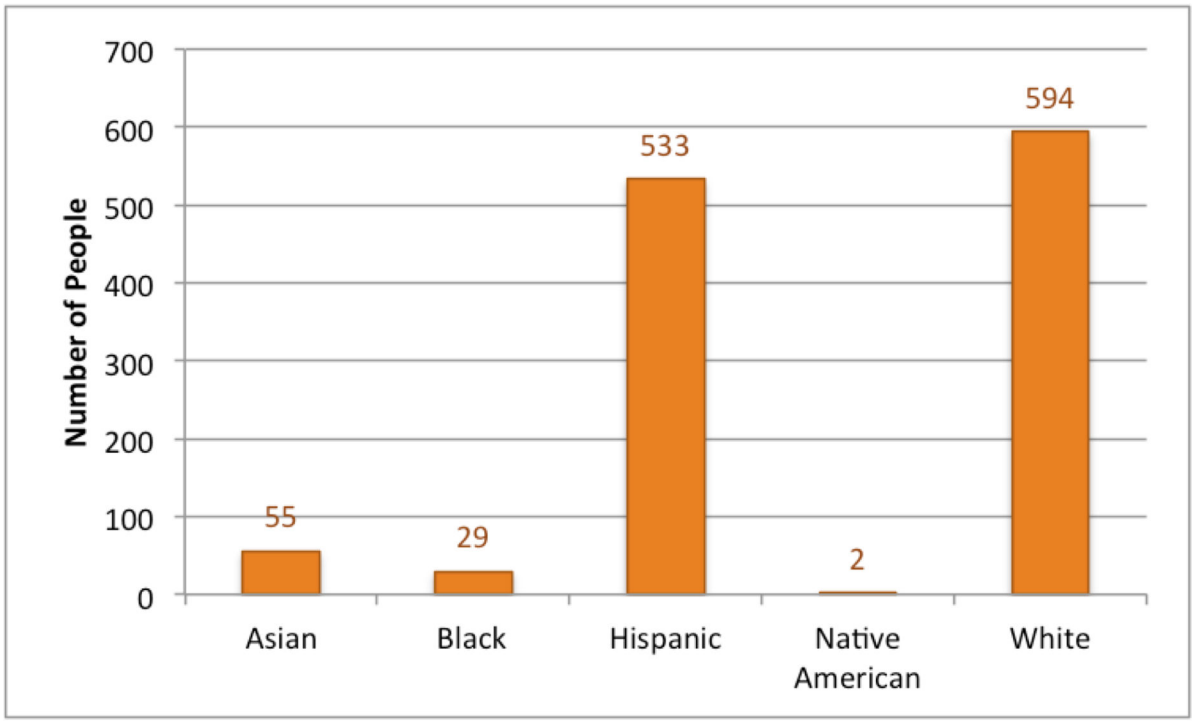
(a)



(b)



(c)

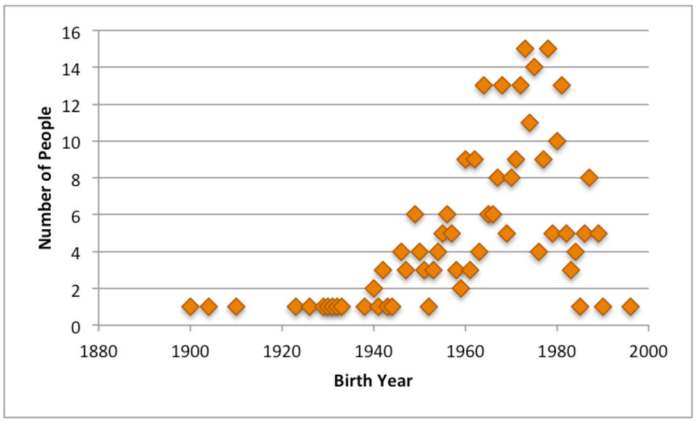


(d)

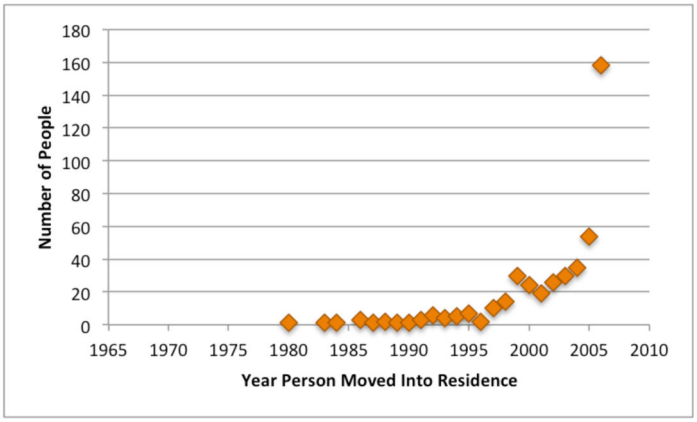
Figure 12.

Distributions of information about named people who lived in Atchison Village in 2006: (a) the number of people per address; (b) their birth years; (c) the year in which they moved into their residence; and, (d) residents' race or ethnicity inferred from last name by computer program. Using last names to infer race is not a good predictor of Blacks; some number of those identified as white may be black.

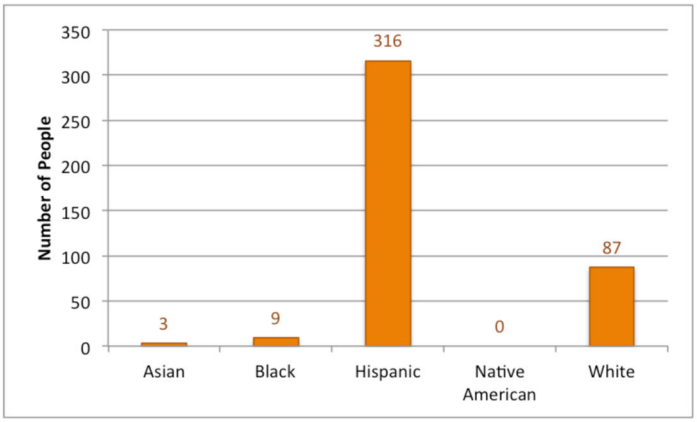
(a)



(b)



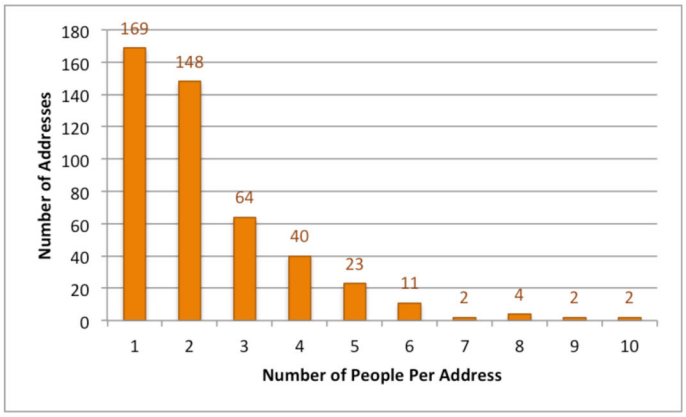
(c)



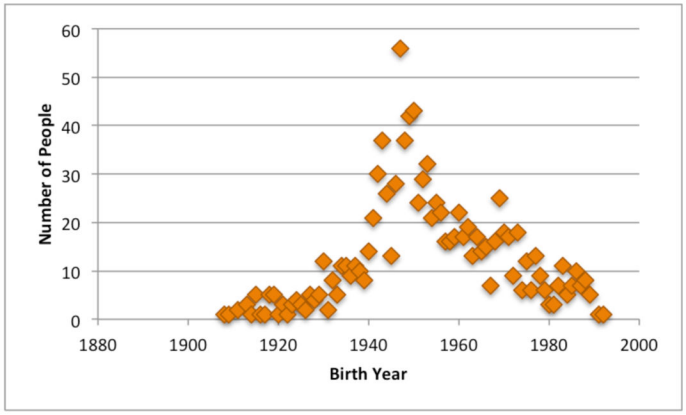
(d)

Figure 13.

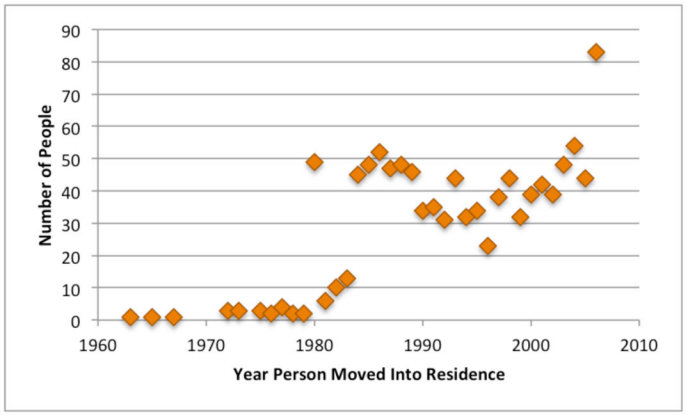
Distributions of information about named people who lived in Liberty Village in 2006: (a) the number of people per address; (b) their birth years; (c) the year in which they moved into their residence; and, (d) residents' race or ethnicity inferred from last name by computer program. Using last names to infer race is not a good predictor of Blacks; some number of those identified as white may be black.



(a)



(b)



(c)

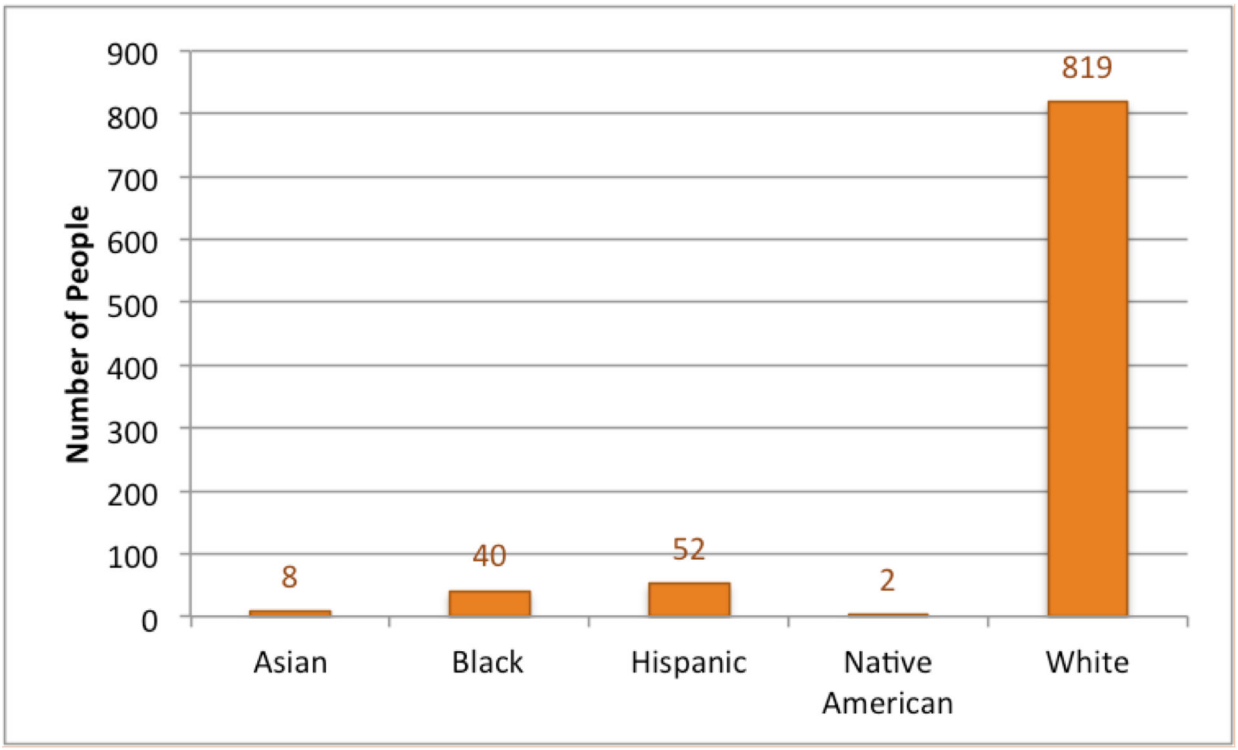


Figure 14.

Distributions of information about named people who lived in Bolinas in 2006: (a) the number of people per address; (b) their birth years; (c) the year in which they moved into their residence; and, (d) residents' race or ethnicity inferred from last name by computer program. Using last names to infer race is not a good predictor of Blacks; some number of those identified as white may be black.

Number of Property Records in a Match (Binsize)	Cumulative Number of Matches
1	2
3	3
5	4
7	6
48	7
68	11
96	20
215	29
223	32

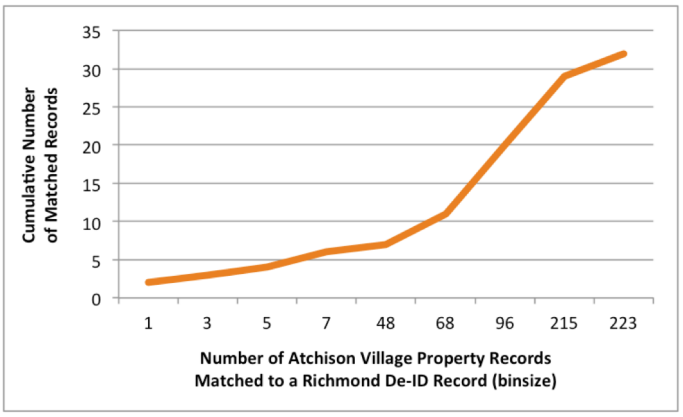


Figure 15.

Matches of records in the Atchison Village Property Register to 32 records in the De-ID Dataset based on housing characteristics alone. The binsize is the number of property records matched to the same De-ID Dataset record. Two records matched uniquely. Three records in the De-ID Dataset each matched to 223 records in the Atchison Village Property Register. The same record in the Atchison Village Property Register may match to more than one record in the De-ID Dataset.

Number of Named People Matched to Same De-ID Record (Binsize)	Number of Binsized Groups
1	5
2	1
3	1
6	1
7	1
8	2
10	1
11	1
14	2
15	1
18	1

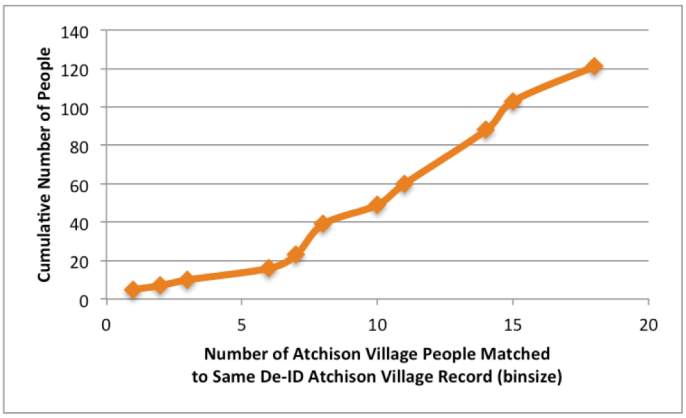


Figure 16.

Small group re-identification (k<20) of named Atchison Village people based on housing characteristics and personal demographics with human-assigned values for gender and race. Matching records in the combined Atchison Village property and De-ID Dataset were further matched to records in the Atchison Village People Register; see Figure 1(b). Binsize is the number of named people matched to combined De-ID and property records. A total of 121 matches to named people appeared for 17 of 32 (or 53 percent) of the De-ID Dataset records for Atchison Village having small group (k<20) re-identifications.

Number of Addresses Matched to Same De-ID Record (Binsize)	Number of Binsized Groups
1	5
2	1
3	1
6	1
7	3
10	1
11	1
14	3
16	1
19	1

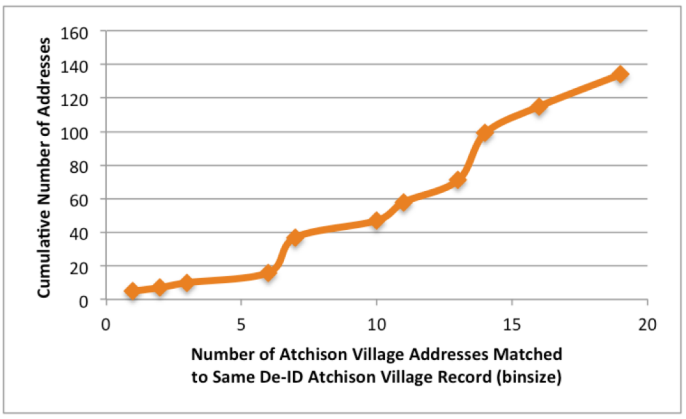


Figure 17.

Small group re-identifications (k<20) of Atchison Village addresses based on housing characteristics and personal demographics with human-assigned values for gender and race. Matching records in the combined Atchison Village property and De-ID Dataset were further matched to records in the Atchison Village People Register; see Figure 1(b). Binsize is the number of addresses matched to combined De-ID and property records. A total of 135 small group (k<20) matches to addresses appeared for 18 of 32 (56 percent) of the De-ID Dataset records for Atchison Village.

Number of Named People Matched to Same De-ID Record (Binsize)	Number of Binsized Groups
1	2
3	3
4	2
5	1
6	1
7	2
8	2
10	1
11	1
12	3
14	2
17	1

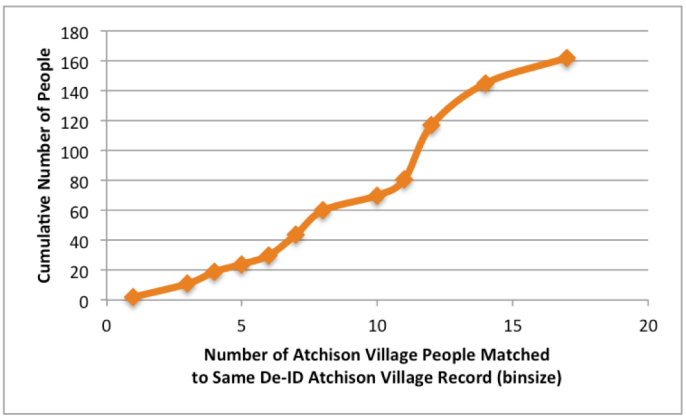


Figure 18.

Small group re-identifications (k<20) of named Atchison Village people based on housing characteristics and personal demographics with computer-assigned values for gender and race. Matching records in the combined Atchison Village property and De-ID Dataset were further matched to records in the Atchison Village People Register; see Figure 1(b). Binsize is the number of named people matched to the combined De-ID and property records. A total of 162 small group (k<20) matches of named people appeared for 21 of 32 (66 percent) of the De-ID Dataset records for Atchison Village.

Number of Addresses Matched to Same De-ID Record (Binsize)	Number of Binsized Groups
1	2
3	3
4	2
5	2
7	2
8	2
10	1
11	3
12	1
14	2
17	1

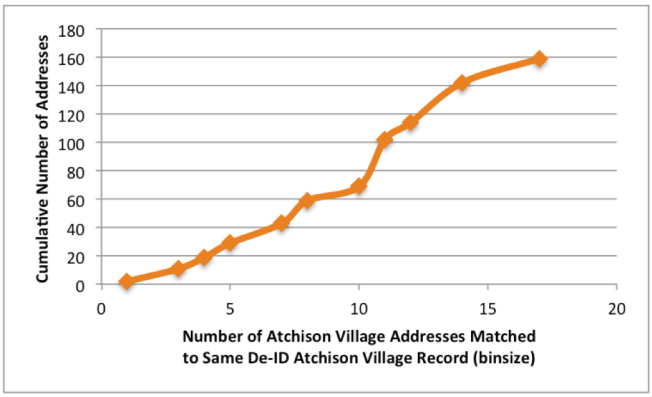


Figure 19.

Small group re-identifications (k<20) of Atchison Village addresses based on housing characteristics and personal demographics with computer-assigned values for gender and race. Matching records in the combined Atchison Village property and De-ID Dataset were further matched to records in the Atchison Village People Register; see Figure 1(b). Binsize is the number of addresses matched to combined De-ID and property records. A total of 159 matches of addresses appeared for 21 of 32 (66 percent) of the De-ID Dataset records for Atchison Village having small group (k<20) re-identifications.

Atchison Village Re-Identification Strategy	k=1	k<5		k<11		k<20	
		Pool size	# of groups	Pool size	# of groups	Pool size	# of groups
People: Hand-Label Gender, Race	5	9	7	40	12	109	17
Address: Hand-Label Gender, Race	5	9	7	39	12	109	18
People: Computer-Assign Gender, Race	2	15	7	58	14	124	21
Address: Computer-Assign Gender, Race	2	16	7	57	14	108	21

Figure 20.

Summary of Atchison Village re-identification pools and number of re-identification groups for named people and addresses using people registers having hand labeled and computer labeled values for gender and race/ethnicity. The size of the re-identification pool and the number of re-identification groups appear for binsizes of k=1 (unique re-identifications), k<5, k<11, and k<20.

Atchison Village People Re-identification Scores					
Hand-Label Gender, Race			Computer-Assign Gender, Race		
Group size	Total no. of groups	No. of groups having a correct match name	Group size	Total no. of groups	No. of groups having a correct match name
1	5	0	1	2	1
2	1	0	3	3	2
3	1	1	4	2	1
6	1	1	5	1	0
7	1	0	6	1	1
8	2	1	7	2	2
10	1	1	8	2	2
11	1	0	10	1	1
14	2	2	11	1	1
15	1	1	12	3	2
18	1	0	14	2	2
TOTAL	17	7	17	1	1
			TOTAL	21	16

Figure 21.

Scored results for Atchison Village people re-identifications for binsizes less than 20. Left side reports results for data having manually labeled gender and race, in which 7 of 17 or 41 percent of the groups included the correct person. Right side reports results for data having computer-assigned labels for gender and race, in which 16 of 21 or 76 percent of the groups included the correct address.

Atchison Village Address Re-identification Scores					
Hand-Label Gender, Race			Computer-Assign Gender, Race		
Group size	Total no. of groups	No. of groups having a correct match address	Group size	Total no. of groups	No. of groups having a correct match address
1	5	0	1	2	1
2	1	0	3	3	2
3	1	1	4	2	1
6	1	1	5	2	1
7	3	2	7	2	2
10	1	1	8	2	2
11	1	0	10	1	1
14	3	3	11	3	2
16	1	1	12	1	1
19	1	1	14	2	2
TOTAL	18	10	17	1	1
			TOTAL	21	16

Figure 22.

Scored results for Atchison Village address re-identifications for binsizes less than 20. Left side reports results for data having manually labeled gender and race, in which 10 of 18 or 56 percent of the groups included the correct address. Right side reports results for data having computer-assigned labels for gender and race, in which 16 of 21 or 76 percent of the groups included the correct address.

Number of Property Records in a Match (binsize)	Cumulative Number of Matches
29	2
75	3
83	9

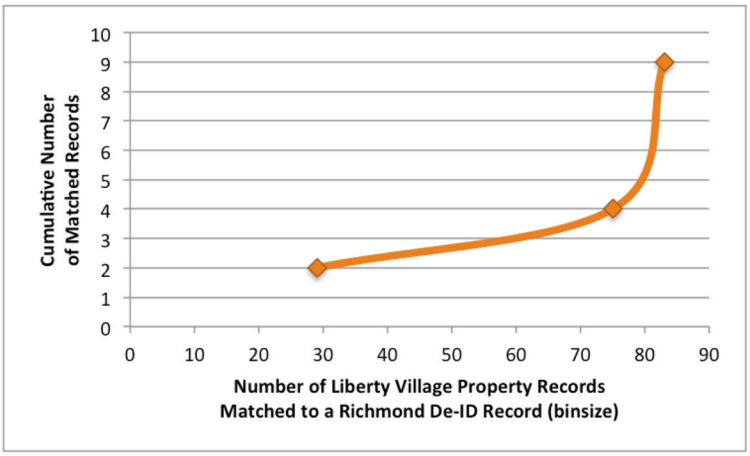


Figure 23.

Matches of records in the Liberty Village Property Register to records in the De-ID Dataset based on housing characteristics alone. The binsize is the number of property records matched to the same De-ID Dataset record. There were no unique or small group matches. The same record in the Liberty Village Property Register may match to more than one record in the De-ID Dataset.

Liberty Village People Re-identification Scores					
Hand-Label Gender, Race			Hand-Label Gender, Race		
Group size	Total no. of groups	No. of groups having a correct match name	Group size	Total no. of groups	No. of groups having a correct match address
1	1	0	1	1	0
7	3	0	6	3	1
18	3	1	17	3	1
TOTAL	7	1	TOTAL	7	2

Figure 24.

Scored results for Liberty Village People Re-identifications for binsizes of 20 or less. The data have manually labeled gender and race. One of 7 of the groups included the correct person by name (left), and 2 of 7 of the groups included the correct address (right).

Number of Property Records in a Match (binsize)	Cumulative Number of Matches
4	2
9	4
12	5
21	6
24	8
28	9
65	10

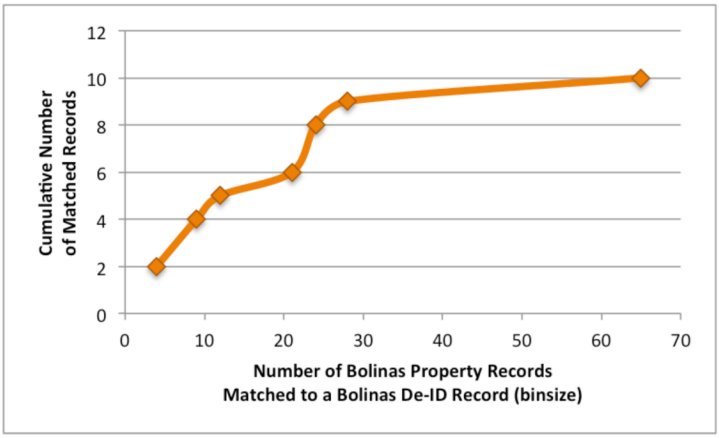


Figure 25.

Matches of records in the Bolinas Property Register to records in the De-ID Dataset based on housing characteristics alone. The binsize is the number of Bolinas property records matched to the same Bolinas De-ID Dataset record. No records matched uniquely. A total of 163 Bolinas addresses matched to the 10 Bolinas De-ID records.

Atchison Village People Re-identification Scores					
Year of Birth			Year of Birth		
Group size	Total no. of groups	No. of groups having a correct match name	Group size	Total no. of groups	No. of groups having a correct match address
1	11	8	1	11	9
2	5	5	2	5	5
3	2	2	3	2	2
TOTAL	18	15	TOTAL	18	16

Figure 26.

Scored results for Atchison Village People Re-identifications for binsizes of 20 or less using year of birth information. Left side reports results for named people, in which 15 of 18 or 83 percent of the groups included the correct person. Right side reports results for addresses, in which 16 of 18 or 89 percent of the groups included the correct address.