PRIVACYANALYTICS
Data Anonymization Solutions
White paper

De-Identification 101

We live in a world today where our personal information is continuously being captured in a multitude of electronic databases. Details about our health, financial status and buying habits are stored in massive databases managed by public and private sector organizations. The “Big Data” age is here and has presented organizations with new opportunities and risks. These databases contain information about thousands of people and can provide valuable research, epidemiological and business insights. This Privacy Analytics white paper looks at the how we can unlock this valuable data while protecting individual privacy.

The Risks of Disclosing Personal Data

To extract or maximize the value contained in databases, data custodians must often provide outside organizations access to their data. In order to protect the privacy of the individuals whose data is being disclosed, a data custodian must “de-identify” information before releasing it to a third-party. De-identification ensures that data cannot be matched to the person it describes. What might seem like a simple matter of masking a person’s direct identifiers (name, address), the problem of de-identification has proven more difficult and is an active area of scientific research.

Who is Affected by the Requirement for De-Identification?

Many governments have enacted legislation requiring organizations to adopt measures to protect personal data. For example, in the United States, health information is protected by the Health Insurance Portability and Accountability Act (HIPAA) and financial information by the Sarbanes-Oxley Act (SOX). Similar legislation exists in the European Union and Canada. The problem of de-identification affects a variety of industries for a multitude of purposes including:

Research

E.G. Registries. Health care organizations (e.g., hospitals, clinics) currently submit patient data to registries. Data contained in these registries can be used for research and policy/administrative needs (such as a stroke or cancer registry). Often data is disclosed from a registry without patient consent under the assumption that it is de-identified.

Open Data

E.G. Government and National Statistical Agencies. A census agency is the most commonly known provider of de-identified information. Census results are de-identified and provided/sold to third parties for further analysis. Open data initiatives are focused on unleashing the potential of the data for the creation of innovative products and services, for creating transparency, to increase service offerings to citizens or to allow citizens to have more control over their health-care. For example, the U.S. Government has developed the “Digital Government Strategy” to build a “21st Century Platform to Better Serve the American People.”[1] Or, consider the State of Louisiana Department of Health and Hospitals example. They are utilizing data to raise the state’s rankings in America’s Health Rankings. De-identified data was made available for an open competition to leverage innovative technologies to help citizens of the State “Own Their Own Health”[12]

Software Testing

E.G. Healthcare IT developers. In the instance where an organization is developing or maintaining health information systems or operations, there is the need to provide developers/QA teams with test data. Often, personal data is taken from a production system and must then be de-identified before being provided to the testing team.

Drug Alerts

E.G. Pharmaceutical firms. Data brokers currently collect prescription data and sell the analysis derived from it to pharmaceutical companies. Personal information must be de-identified before being sent to a data broker.

Data Warehouses

E.G. Insurance.Like pharmaceutical companies, insurance companies analyze claims data for actuarial and marketing reasons. De-identification is required to comply with privacy best practices, and in some jurisdictions, regulations.

Medical Devices

E.G. Dialysis machine manufacturers. Medical device companies receive data from the devices they manufacture. These types of devices include dialysis machines, heart monitors, MRIs etc. The data can then be de-identified by the medical device company and used for analytics purposes (eg. diagnosis and trend analysis).

What are the Motivations to Protect the Privacy of Individuals?

Litigation

Depending on the jurisdiction of the incident, if a person’s private information is released by an organization without the person’s consent, that person has the right to file a complaint with a regulatory authority or take the organization to court. This can lead to a costly settlement or to litigation, even if no damages are awarded.

Cost

If an organization inadvertently releases private information, privacy legislation often mandates that the people whose data was exposed must be notified. In addition to the cost of breach notification, an organization might face significant compensation costs, and increasingly, fines by regulators.

Reputation

A privacy breach is a public relations disaster for an organization (public or private), and can directly affect the bottom line. Furthermore, breaches erode the public/client/patient trust in that organization.

Examples of Re-Identification

To avoid privacy breaches, organizations currently use manual, ad-hoc methods to de-identify personal information. Given the lack of publically available de-identification tools that have been proven to be effective, there have been several high-profile incidents where improper de-identification has resulted in a privacy breach. Recent examples include:

I) Data from the Group Insurance Commission, which purchases health insurance for state employees in Massachusetts, was matched against the voter list for Cambridge, re-identifying the governor’s record.

II) Students were able to re-identify a significant percentage of individuals in the Chicago homicide database by linking with the social security death index.

III) Individuals in a de-identified/anonymized publicly available database of customer movie recommendations from Netflix were re-identified by linking their ratings with ratings in a publicly available Internet movie rating web site.

IV) A national broadcaster aired a report on the death of a 26 year-old female taking a particular drug who was re-identified from the adverse drug reaction database released by Health Canada.

V) AOL put de-identified/anonymized Internet search data (including health-related searches) on its web site. New York Times reporters were able to re-identify an individual from her search records within a few days.

These re-identifications were possible because the methods for de-identification utilized were not effective or conducted in a defensible way and did not ensure that the risk of re-identification was sufficiently low before the data was disclosed. Proper de-identification would have made those breaches highly unlikely.

What are the Standards for De-identification?

One of the main standards used as guidance for de-identifying personally identifiable information (PII) and personal health information(PHI) is the HIPAA Privacy Rule(45 CFR 164.514) from the US Department of Health and Human Services. It was designed to protect personally identifiable health information through permitting only certain uses and disclosures of PHI provided by the Rule, or as authorized by the individual subject of the information.[2]

The HIPAA Privacy Rule provides mechanisms for using and disclosing health data responsibly without the need for patient authorization. These mechanisms center on the HIPAA de-identification standards – HIPAA Safe Harbor and the Statistical or Expert Determination methods.

Safe Harbor Direct and Quasi Identifiers

1) Names
2) Zip Codes (except first three)
3) All elements of dates (except year)
4) Telephone Numbers
5) Fax Numbers
6) Electronic Mail Addresses
7) Social Security Numbers
8) Medical Record Numbers
9) Health Plan Beneficiary Numbers
10) Account Numbers
11) Certificate/License Numbers
12) Vehicle Identifiers and Serial Numbers, including license plate numbers
13) Device Identifiers and Serial Numbers
14) Web Universal Resource Locators (URL)
15) Internet Protocol (IP) Address Numbers
16) Biometric Identifiers, including finger and voice prints
17) Full face photographic images and any comparable images
18) Any other unique identifying number, characteristic or code

The Two Methods of De-identification are:

HIPAA Safe Harbor Method

Expert Determination/Statistical Method

HIPAA Safe Harbor Method

I. Removal of 18 types of direct and quasi-identifiers

II. No actual knowledge residual information can identify an individual

Expert Determination Method/Statistical Method

A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

I. Applying such principles and methods, determines that the risk is **very small** that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and

II. Documents the methods and results of the analysis that justify such determination.

It is not possible to have zero risk with either of the two de-identification methods defined. However, it is possible to have very small risk with the Statistical Method. The possibility does exist that the de-identified data could be linked back to the patient. Regardless of the method by which de-identification is achieved, the Privacy Rule does not restrict the use or disclosure of de-identified health information, as it is no longer considered protected health information.[2]

A note of importance is that HIPAA applies to Covered Entities, which are health plans, healthcare providers, and data clearinghouses. Many organizations that wish to share health data may not fall under HIPAA, but should consider adhering to this standard as a means of good practice. Organizations should do their homework early on to determine if they fit into the above category or are classified as a Business Associate (BA). The January 25, 2013 Omnibus Rule for HIPAA implemented statutory amendments under the Health Information Technology for Economic and Clinical Health Act (HITECH) with regards to Business Associates. In particular, is a significant change to the liability for Business Associates. In the Omnibus Rule, HHS has increased the liability for Business Associates and now makes them directly liable for:

- Impermissible uses and disclosures
- Failure to provide breach notification to the covered entity;
- Failure to provide access to a copy of electronic PHI to either the covered entity, the individual, or the individual's designee (whichever is specified in the business associate agreement);
- Failure to disclose PHI when required in an investigation of the BA's compliance with HIPAA;
- Failure to describe when an individual's information is disclosed to others; and

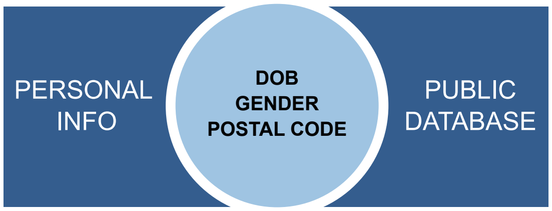
- Failure to comply with the HIPAA Security Rule’s requirements, such as performing a risk analysis, establishing a risk management program, and designating a security official, among other administrative, physical, and technical safeguards.[5]

Under the final rule, BA’s will face civil monetary penalties that range from $100 to $50,000 per violation, with a cap of $1.5 million for multiple violations of the same provision. BA’s will need to ensure they are in compliance with the final rule by September 23, 2013. For confirmation on whether or not your organization is a covered entity or not, the Centers for Medicare and Medicaid Services provides an easy-to-use question and answer decision tool3 to help you decide. To determine if you are considered a BA under HIPAA, the Department of Health and Human Services provides definitions, background and examples for your reference.6 Examples of Business Associates include third party administrators/claims processors for health plans, attorneys that have access to their clients PHI or a third party researcher.

In light of the facts that the Safe Harbor Method of de-identification really only provides a “check in the box” for HIPAA compliance, that PHI that has been de-identified, does not yield high utility data for use or disclosure for secondary purposes, and it increases the risks of leakage of sensitive information when that “de-identified” data is mixed with other data sets for analysis; we recommend that covered entities and business associates should use the Expert Determination/ Statistical Method of de-identification to ensure they are compliant with the HIPAA Privacy Rule.

Quasi-Identifiers: The Devil is in the Details

Many people assume that by removing names and addresses (direct identifiers) when de-identifying records that it is sufficient to protect the privacy of the persons whose data is being released. The problem with comprehensive de-identification is that it also involves those personal details that are not obviously identifying. These personal details, known as quasi-identifiers, include the person’s age, sex, postal code, profession, ethnic origin and income (to name a few).



Privacy Analytics Inc. has integrated algorithms developed by the Electronic Health Information Laboratory (EHIL) into the Privacy Analytics Risk Assessment Tool (PARAT) to allow organizations to measure re-identification risk and de-identify their data.

EHIL has focused its research on the de-identification of quasi-identifiers. The three unique types of re-identification attacks highlighted are: prosecutor, journalist, and marketer. Algorithms developed by the lab measure the risk of each type of attack. In addition to rigorous testing, the work has been published in peer-reviewed journals (see the Publications section for details).

Prosecutor Risk

In this scenario, an intruder wants to re-identify a specific person in a de-identified database. Let's take the example of an employer who has obtained a de-identified database of drug test results. The employer is trying to find the test results of one of their employees (Dave, a 37 year-old doctor) and knows that Dave's record is in the de-identified dataset.

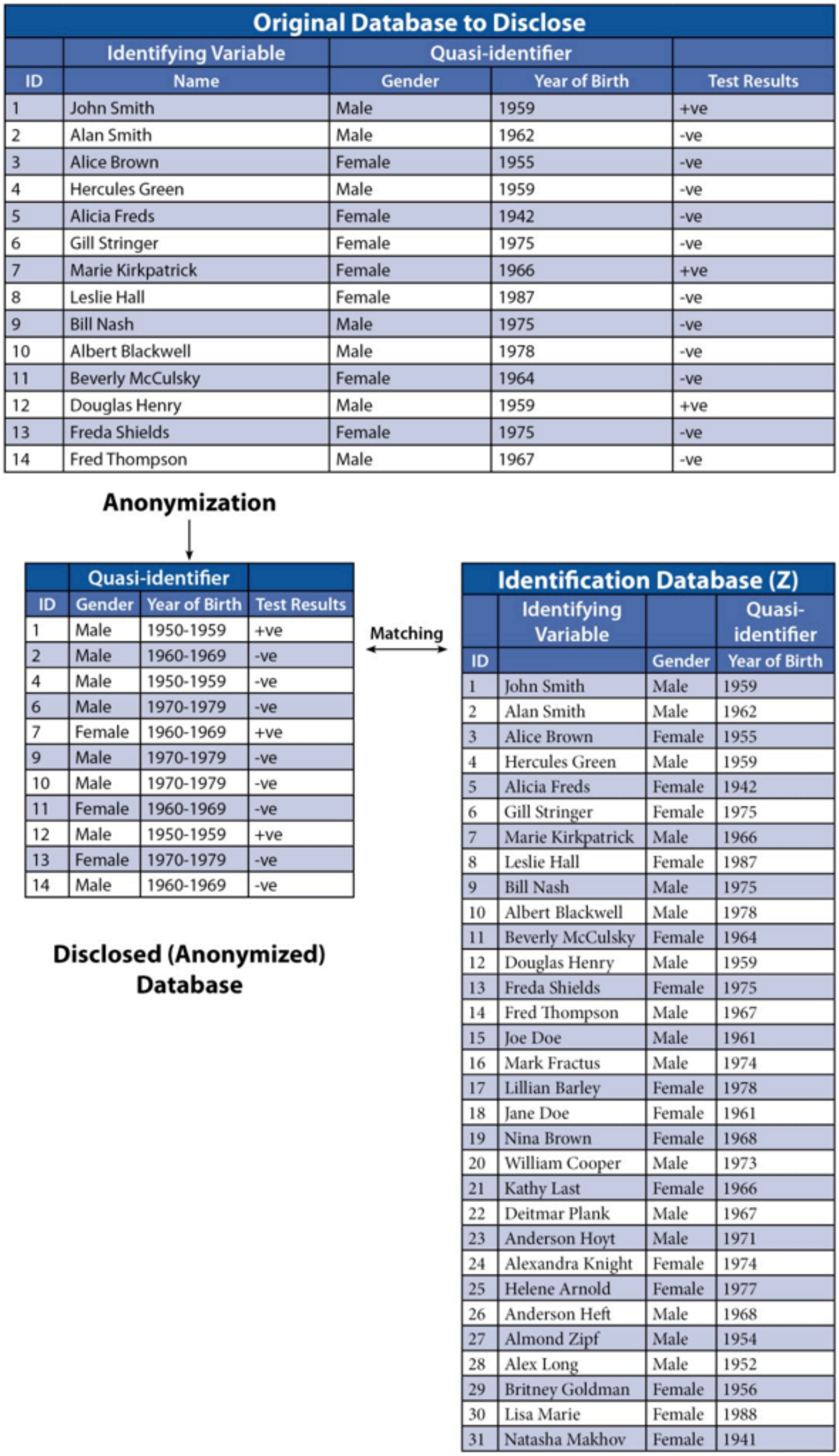
The re-identification risk is measured by finding the unique combinations of quasi-identifiers in the de-identified/anonymized data set (these are called equivalence classes). To illustrate what is an equivalence class, let's take the following data set containing the quasi-identifiers of sex, age and profession. The data set also contains the person's latest drug test results (this is the sensitive data).

In this data set there are three equivalence classes: 39 year-old male doctors, 37-year-old male doctors and 39-year old female doctors. Since the employer knows that Dave is a 37 year-old doctor, there is a 1 in 3 chance (33%) of identifying Dave's record correctly. If however, the employer were attempting to identify a 39-year old female doctor, there would be a perfect match since only one record in that equivalence class exists. Since we cannot predict which equivalence class an intruder will attempt to match, we must assume the worst-case scenario, which is that the person they want to identify has the smallest equivalence class (denoted by k) in the database (i.e., 39-year-old female doctor). When de-identifying a data set, a value of 5 for k (i.e., there are at least five records in any equivalence class) is often considered sufficient privacy protection.

ID	Sex	Age	Profession	Drug Test
1	Male	37	Doctor	Negative
2	Female	39	Doctor	Positive
3	Male	37	Doctor	Negative
4	Male	39	Doctor	Positive
5	Male	39	Doctor	Negative
6	Male	37	Doctor	Negative

Journalist Risk

Journalist risk is also concerned with the re-identification of individuals. However, in this case the journalist does not care which individual is re-identified. The probabilistic risk profile here is quite different from that of prosecutor risk. In the journalist scenario, the de-identified/anonymized data is a subset of a larger public database. The journalist doesn't know a particular individual in the anonymized data set but does know that all the people in the data set exist in a larger public database (which they have access to). A real-life example of a journalist attack occurred when a Canadian Broadcasting Corporation (CBC) reporter re-identified a patient in a de-identified adverse drug reaction database by matching her age, date of death, gender, and location with the public obituaries. Previous research has shown that the smallest equivalence class found in the public database that maps to the anonymized data set measures the risk of re-identification. To illustrate this, let's look at the following tables.



Original Database to Disclose

	Identifying Variable	Quasi-identifier		
ID	Name	Gender	Year of Birth	Test Results
1	John Smith	Male	1959	+ve
2	Alan Smith	Male	1962	-ve
3	Alice Brown	Female	1955	-ve
4	Hercules Green	Male	1959	-ve
5	Alicia Freds	Female	1942	-ve
6	Gill Stringer	Female	1975	-ve
7	Marie Kirkpatrick	Female	1966	+ve
8	Leslie Hall	Female	1987	-ve
9	Bill Nash	Male	1975	-ve
10	Albert Blackwell	Male	1978	-ve
11	Beverly McCulsky	Female	1964	-ve
12	Douglas Henry	Male	1959	+ve
13	Freda Shields	Female	1975	-ve
14	Fred Thompson	Male	1967	-ve

Anonymization

	Quasi-identifier		
ID	Gender	Year of Birth	Test Results
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Male	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
14	Male	1960-1969	-ve

Disclosed (Anonymized) Database

Matching

Identification Database (Z)

	Identifying Variable		Quasi-identifier
ID		Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Male	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Male	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Makhov	Female	1941

The first table is the original data set before anonymization. The records in the table are a subset of those found in registry (Z). The data set is anonymized by removing names and aggregating the year of birth by decade (decade of birth). There are five equivalence classes in the anonymized table that map to the public registry which can be found in this table.

Equivalence Class		**Registry Table**		**Public Registry**	
Gender	Age	Count	ID	Count	ID
Male	1950-1959	3	1,4,12	4	1,4,12,27
Male	1960-1969	2	2,14	5	2,14,15,22,26
Male	1970-1979	2	9,10	5	9,10,16,20,23
Female	1960-1969	2	7,11	5	7,11,18,19,21
Female	1970-1979	2	6,13	5	6,13,17,24,25

This table shows that the smallest equivalence class in the public database (Z) that map to the anonymized data set is a male born in the 1950s (four records). Therefore, there is a one in four chance (25%) of re-identifying a record that falls in this equivalence class. The problem with applying the existing journalist re-identification risk analysis is that the entire content of the public database (Z) is rarely known (e.g., due to cost, logistics, legal, retension). To overcome this limitation, the researchers at EHIL developed a method to estimate the number of records found in each equivalence class in a public registry. This allows the re-identification risk in the journalist scenario to be calculated and controlled without having access to the larger public database.

Marketer risk

In this scenario, an intruder wants to re-identify as many individuals as possible in a database. The marketer is less concerned if some of the records are misidentified. Therefore, rather than focus on individuals, here the risk pertains to everyone in the data set. Take for example a pharmaceutical company that obtained de-identified prescription data. They can attempt to match this data with their internal marketing database to create a mailing campaign (say, targeting doctors or patients). They are not concerned if some of the mailers are sent to the wrong physicians (i.e., spam).

The marketer risk is measured by calculating the probability of matching a record in an equivalence class of the de-identified set with those in the matching equivalence class in the marketer's database. In the journalist example (see above), the first equivalence class (males ages 1950-1959) has three records that could be matched to one of four possible records in the public registry. The expected number of records that a marketer can properly identify when randomly matching records in the de-identified data set with those in the public database can be calculated for each equivalence class.

Equivalence Class		Anonymized Table		Public Registry		Expected # Correct Matches
Gender	Age	Count	ID	Count	ID	
Male	1950-1959	3	1,4,12	4	1,4,12,27	3 / 4
Male	1960-1969	2	2,14	5	2,14,15,,22,26	2 / 5
Male	1970-1979	2	9,10	5	9,10,16,20,23	2 / 5
Female	1960-1969	2	7,11	5	7,11,18,19,21	2 / 5
Female	1970-1979	2	6,13	5	6,13,17,24,25	2 / 5
Expected number of identified records						2.35

A marketer would expect to correctly re-identify about 21% (2.35/11) of the overall records in this scenario.

De-Identifying Data

Besides the standards for de-identification, there are several options available to an organization on how to go about de-identifying its data. Organizations can employ in-house homegrown solutions that typically apply HIPAA Safe Harbor. They can engage de-identification consultants that are qualified to de-identify data under HIPAA and certify that the data set is defensible and provide an audit trail. Or finally, they can purchase commercially available software tools and conduct automated in-house de-identification. There exists however, some points of concern with home grown solutions that apply Safe Harbor and de-identification consultant services. For in-house home grown solutions, their methodology may not take into account the risks associated with longitudinal data. They will then find themselves in a situation where the organization is potentially at risk of having data sets that can be re-identified. With regards to de-identification consultants, they will often not want to provide their methodology. In this instance, an organization may not be able to prove that the methodology actually produced a low risk of re-identification which may put them at risk for data breaches.

De-identification techniques include:

Record Suppression

When a record's combination of quasi identifiers presents too high a risk of re-identification to be released, it must be dropped from the data set.

Cell Suppression

A record can be further de-identified by suppressing/masking the value contained in a single field (cell). For example, a field in a patient record containing a very rare disease would be suppressed.

Sub-Sampling

Sub-sampling involves taking a random sample of a data set. For example, if the requirement is to have a data set that is 10% of the original data set, you will get a subset of the original data set that was randomly selected and has 10% of the number of patients as the original.

Aggregation/Generalization

Rare quasi-identifiers can be aggregated to provide better de-identification/anonymization. For example, a low population postal code can be aggregated to a larger geographic area (such as a city). A rare medical profession, such as perinatologist, can be aggregated to a more general obstetrician.

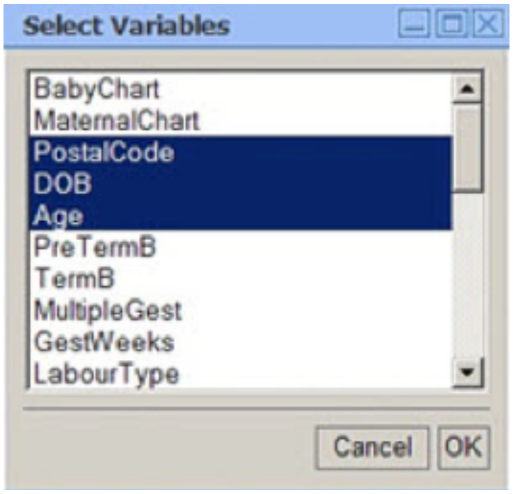
PARAT

PARAT takes the guesswork out of de-identifying personal information. PARAT uses peer-reviewed techniques to measure and manage re-identification risk. Only PARAT can protect against all known types of re-identification attacks. It optimally de-identifies information to protect individual privacy while retaining the data's value. Using a simple four-step process, PARAT allows you to easily and safely release your valuable data.

Step 1: Variable Selection

To begin the process, the quasi-identifiers that are to be released must be selected from the data set.

Once the quasi-identifiers are selected, you can rank them in order of importance (the variables' utility to the person using the de-identified data set). This ranking will be used during the de-identification process to determine the optimal anonymization that balances re-identification risk and data utility. For example, if age is ranked as the most important quasi-identifier and postal code as the least important, the de-identification process will attempt to keep age information intact while the postal code variable will be aggregated (i.e., grouped into larger geographic areas). Ranking allows you to maximize the utility of the de-identified data set.



Step 2: Assign Acceptable Re-Identification Risk Threshold (Safety Index)

PARAT allows you to decide how much de-identification should be done before releasing a data set. The "amount" of de-identification is measured by the probability of accurately re-identifying a record (for prosecutor and journalist risk) or the expected number of records to be re-identified correctly (for marketer risk). For example, if the quasi-identifiers contained in a de-identified record can be associated with five individuals contained a public registry, the probability of re-identification is 0.2 (i.e., 1 in 5 chance of making the correct match).

Achieving a lower probability of re-identification (lower risk) often means reducing the utilityof the released data (either suppressing records or aggregating variables). Ensuring a low re-identification risk might make the de-identified data less useful to the recipient because there is not enough data resolution for their needs.

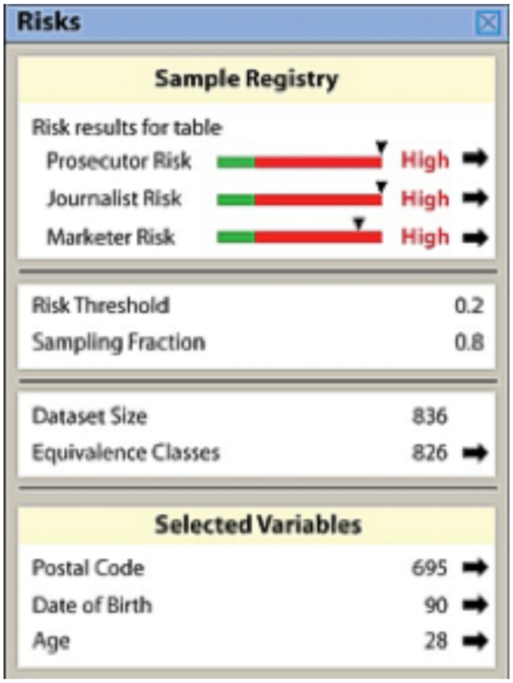
To balance the need for privacy with the need for data resolution, PARAT allows you to set the acceptable probability/risk of re-identification. Re-identification risk can be adjusted based on the profile of the person/organization requesting the information. For example, if data is to be released to the general public, a high degree of de-identification is required (e.g., a threshold of 0.05).

However, if data is being shared within an organization (e.g., between government departments), a lesser amount of de-identification is needed (e.g., a threshold of 0.2). To help determine what is the right amount of de-identification, we provide a methodology to rate the risk of releasing data to a given person or organization. Risk based de-identification ensures that individual privacy is protected while optimizing the released data's value.

Step 3: Risk Measurement

Once the acceptable threshold has been set, the risk analysis can be performed. PARAT calculates the data set's risk for the three types of re-identification attacks: prosecutor, journalist and marketer.

In this example, a data set containing the quasi-identifiers of postal code, date of birth and age has been analyzed with a re-identification risk threshold of 0.2. The results show the re-identification risk is high (above 0.2) for all three types of attacks: prosecutor, journalist, and marketer. Of the 836 records in the data set, 826 have a unique combination of quasi-identifiers (equivalence classes). The data set contains 695 unique postal codes, 90 unique birth dates and 28 unique ages.



Risks

Sample Registry

Risk results for table

Prosecutor Risk	High	→
Journalist Risk	High	→
Marketer Risk	High	→

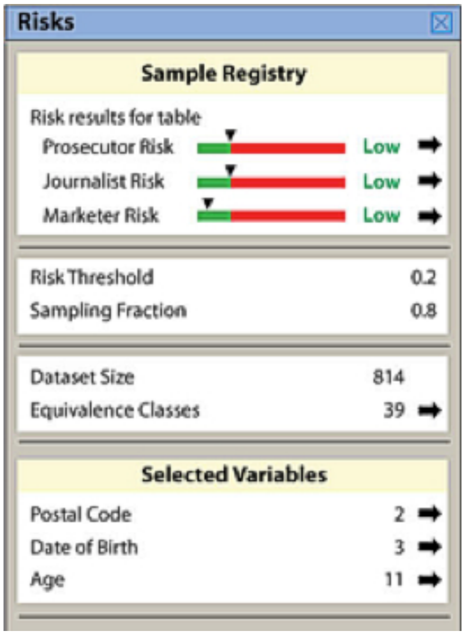
Risk Threshold	0.2	
Sampling Fraction	0.8	
Dataset Size	836	
Equivalence Classes	826	→

Selected Variables

Postal Code	695	→
Date of Birth	90	→
Age	28	→

Step 4: De-Identification

To reduce the risk of re-identification below our acceptable threshold (0.2 in this example), PARAT will optimally de-identify the data. After the de-identification process, the risk for all types of re-identification attacks has been reduced to acceptable levels. This was done by marking 22 records for suppression and aggregating quasi-identifier values. Postal code values are grouped into two areas; dates of birth are aggregated into three ranges and age into 11 ranges.



Risks

Sample Registry

Risk results for table
Prosecutor Risk Low
Journalist Risk Low
Marketer Risk Low

Risk Threshold 0.2
Sampling Fraction 0.8

Dataset Size 814
Equivalence Classes 39

Selected Variables

Postal Code 2
Date of Birth 3
Age 11

PARAT automatically produces the optimally de-identified/anonymized data set that meets the desired re-identification risk threshold.

Age Before De-Identification

Count	Age
3	43
5	16
6	18
6	17
7	42
8	19
9	20
9	41
14	40
14	21
20	23
24	27
25	39
28	38
30	22
33	24
33	26
34	25
41	27
45	31
46	28

Age After De-Identification

Count	Age
19	41-45
34	16-20
131	21-25
140	36-40
247	26-30
264	31-35

Publications

F. K. Dankar and K. El Emam: “A method for evaluating marketer re-identification risk”. In Proceedings of the 3rd International Workshop on Privacy and Anonymity in the Information Society, 2010.

K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, JP. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, J. Bottomley: “A Globally Optimal k-Anonymity Method for the Deidentification of Health Data .” In Journal of the American Medical Informatics Association, 16(5):670-682, 2009.

K. El Emam, A. Brown, and P. AbdelMalik: “Evaluating predictors of geographic area population size cutoffs to manage re-identification risk.” In Journal of the American Medical Informatics Association, March/April, 16(2):256-266, 2009.

K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk: “Evaluating Patient Re-identification Risk from Hospital Prescription Records.” In the Canadian Journal of Hospital Pharmacy, 62(4):307-319, 2009.

K. El Emam: “Heuristics for de-identifying health data.” In IEEE Security and Privacy, July/August, 6(4):58-61, 2008.

K. El Emam, and F. Dankar: “Protecting privacy using k-anonymity.” In the Journal of the American Medical Informatics Association, September/October, 15:627-637, 2008.

K. El Emam, E. Neri, and E. Jonker: “An evaluation of personal health information remnants in second hand personal computer disk drives.” In Journal of Medical Internet Research, 9(3):e24, 2007.

K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, M. Power: “Evaluating common de-identification heuristics for personal health information.” In Journal of Medical Internet Research, 2006;8(4):e28, November 2006.

K. El Emam: “Overview of Factors Affecting the Risk of Re-Identification in Canada”, Access to Information and Privacy, Health Canada, May 2006.

K. El Emam: “Data Anonymization Practices in Clinical Research: A Descriptive Study”, Access to Information and Privacy, Health Canada, May 2006.

References

1. The White House, "Digital Government – Building a 21st Century Platform to Better Serve the American People," Accessed March 2013

2. State of Louisiana, Department of Health and Hospitals, "Taking the Living Well in Louisiana Challenge", http://new.dhh.louisiana.gov/index.cfm/page/1327, Accessed May 2013

3. Department of Health and Human Services, Office for Civil Rights, "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#standard, Accessed February 2013

4. Centers for Medicare and Medicaid Services, "Are you a Covered Entity", http://www.cms.gov/Regulations-and-Guidance/HIPAA-Administrative-Simplification/HIPAAGenInfo/AreYouaCoveredEntity.html, Accessed February 2013

5. Brandon C. Ge, "What the HIPAA Omnibus Rule means for health technology companies", http://www.lexology.com/library/detail.aspx?g=92d343de-a2ba-4b5f-826b-e564296ca84d, Accessed March 2013

6. Department of Health and Human Services, Office for Civil Rights, "Business Associates," http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentitities/businessassociates.html, Accessed March 2013

Contact Us

251 Laurier Avenue W, Suite 200
Ottawa, Ontario, Canada
K1P 5J6
Phone: +1.613.369.4313
Toll Free: +1.855.686.4781
Fax: +1.613.369.4312
http://www.privacyanalytics.ca