PharmaSUG 2019 - Paper SS-226

De-Identification of Data & Its Techniques

Shabbir Bookseller, Quartesian.

ABSTRACT

In the past few years, with the emergence of modern technologies in the image of big data, the privacy concerns had grown widely. Now a day's sharing data is much easier then saying Hello. De-identification is a tool that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations. For many reasons, not the least of which is patient privacy, any shared data must first be de-identified De-identification is not a technique of securing data, but a collection of approaches, algorithms, and tools that can be applied to various kinds of data with differing levels of effectiveness of an individual's privacy. In general, privacy protection improves as more aggressive de-identification techniques are employed, but less utility remains in the resulting dataset. De-identification is especially important for government agencies, businesses, and other organizations including healthcare industry that seek to make data available to outsiders. For example, significant medical research resulting in societal benefit is made possible by the sharing of de-identified patient information under the framework established by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, the primary US regulation providing for privacy of medical records. This topic provides an overview of Data De-identification with Data Protection Procedures & its Techniques I.e. K-Anonymity, l-Diversity & t-closeness.

INTRODUCTION - A METHOD TO DISCONNECT IDENTITY FROM THE DATA

Data De-identification is a process of distinguishing private information about a person or entity and removing those identifiers from the data. Extrication of Personally Identifiable Data makes identification of the individual entity, on which the data are based, impossible.

Data De-identification is growing as concept & practice across different spheres of working, especially in healthcare industry wherein medical researchers are increasingly being done on de-identified patient-information under the framework formulated by the Health Insurance Portability & Accountability Act (HIPAA) Privacy Rule, the primary US-regulation to ensure privacy of medical records.

DEEP DIVE INTO DATA-DEIDENTIFICATION

<u>Documents required for de-identification/redaction</u>:

Patient level data and related documents

- SDTM datasets
- Analysis datasets (ADSL)
- Data set specifications or algorithms

Clinical study documents

- Clinical study report (CSR)
- Protocol with any amendments
- Statistical methods within CSR
- Annotated case report form (aCRF)
- Statistical Analysis Plan (SAP)

Procedures:

1. Removing personally identifiable information (PII)-HIPAA 18 identifiers – involves removing of any names, initials, kit numbers, device numbers, geographic info, country level info from variable names e.g. LBNAM containing location information.

2. Recoding identifiers and formal anonymization – European Medicines Agency (EMA) suggest data to be anonymized if personal information is removed or redacted and the SUBJECT ID cannot be linked to a study participant. Therefore, SUBJID are anonymized by replacing the original code number with a new code and destroying the code key that was used to generate the new code number from the original (destroying the link between the two code numbers).

3. Replacing Date of Birth – According to HIPAA rules, DOB is replaced with age in years and all ages above 89 are aggregated into a single category of "90 or older" and for all other ages, range of age is documented. (Exception in Bayer's procedures, BRTHDTC variable greater than 89 or younger are left blank only in the month and day date parts for the date of birth.) All other dates from the other domains like AE, EX, MH, CM and SV will remains when populated.

4. Demographic Information – Aggregating the race = "OTHER" for few patients who falls under a larger race group depending upon the study design. All other DM information which cannot be removed should get advice from the specified biostatistician for further information.

5. SAS formats used to decode INVNAM and SITEID will be removed from the SAS format catalog that accompanies the data.

6. Aggregating center information to the country /region level – SITEID, INVNAM, INVID, COUNTRY will be removed or set to "BLANK" to prevent identification of the location of a subject. COUNTRY with one SITEID will be aggregated to a region, so that location cannot be identified. Likewise, one SITEID specific studies cannot be anonymized and will not available to Data Acquisition System (DAS).

7. Replacing original dates related to a study subject – Replacing RFSTDTC prevents SUBJID in case a date is associated with AESTDTC. RFSTDTC will be shifted by a random, subject specific factor i.e. adding specific offset days to RFSTDTC. Therefore, differences between dates within the subject's data records will be maintained. Two methods can be followed like Dummy date method and Study day method.

   Dummy date method:
   Dates are replaced to new dates by creating a random offset (e.g. 91 days) for each research participant which will be applied to all dates for that research participant and relative times of the dummy days are retained.

	Original Date	New Date	
Reference date	01APR2010	01JUL2010	Apply offset = 91 days
Date of Death	01MAY2010	31JUL2010	Apply offset = 91 days
Relative time of death	30 days	30 days	

   Study day method:
   All dates are removed from the datasets and the study day is calculated for each observation with days relative to a reference date.

	Original date	Reference date	Study Day
Date of Death	01MAY2008	01JAN2008	122

8. Removing comments, free text and free text verbatim terms (CO domains) –All COVALs will removed because it contains patient specific information. AETERM (AE domain) on the CRFs, CMDECOD and CMINDC are removed from CM domains and subject level information from MH domains.

9. Removing all genetic data which track back to an individual subject.

10. Removing of any other uniqueness of Patient record – Aggregating the fields with few patients under a group depending on the study design (e.g. if weight =”224” should become over 200 or between 200 and 300 depending on the weight distribution).

Review Process and Quality Control:

- Final review of HIPAA 18 identifiers is done if further removal is required.
- Quality control review and approval of the anonymized data to confirm no personally identifiable information remains.
- All the applicable changes listed in the documents were applied for redaction.
- Secure storage of the anonymized data.
- Destruction of the code key which links the anonymized dataset and the original dataset.
- Reviewing of free-text analysis variable should be done. Values with PI within the string should be redacted e.g. “Dr. Adam assessed tumor on left arm” redacted to “assessed tumor on left arm”.
- Confirming the no. of records remains constant in a particular original dataset. Reasons must be investigated if there is an alteration due to de-identification.
- Verifying all date variables have been removed and relative study day is included.

The following specific items are discarded (remnants):

- Any transactional copies of anonymized datasets
- De-identification tables
- QC output datasets and review files
- SAS Log and listing files
- Seed utilized for random number generation.
- Exploratory biomarker data outside the primary and key secondary endpoints and laboratory data.
- Case narratives, documentation for adjudication and imaging data (x-rays, MRI scans etc.)

Exclusion of studies which are not possible to anonymize:

High risk sensitive data like studies related to specific mental or sexual diseases, Business Intelligence (BI) assess the anonymization queries and in case of negative outcome will not provide the personally identifiable information (PII) but try to address by providing summary data on requisition.

TECHNIQUES OF DATA DE-IDENTIFICATION: -

- **k-anonymity: A widespread methodology for data anonymization:**

It is a property possessed by certain anonymized data. The concept of k-anonymity was first introduced by Latanya Sweeney and Pierangela Samarati in a paper published in 1998 as an attempt to solve the problem: "Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release. In the context of k-anonymization problems, a database is a table with n rows and m columns. Each row of the table represents a record relating to a specific member of a population and the entries in the various rows need not be unique. The values in the various columns are the values of attributes (variables) associated with the members of the population. See following table as an example for non-anonymized database.

Name	Age	Gender	County	Religion	Disease
John	29	Female	Surrey	Christian	Cancer
Paru	24	Female	Essex	Hindu	Viral infection
Salima	28	Female	Hants	Muslim	TB
Sunny	27	Male	Sussex	Buddhist	No illness
Joan	24	Female	Hants	Christian	Heart-related
Brian	23	Male	Surrey	Christian	TB
Rambo	19	Male	Herts	Hindu	Cancer
Kishor	29	Male	Suffolk	Hindu	Heart-related
Johnson	17	Male	Norfolk	Christian	Heart-related
John	19	Male	Surrey	Christian	Viral infection

There are 6 attributes (variables) and 10 records in this data. There are two common methods for achieving k-anonymity for some value of k.

1. **Suppression**: In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'. In the anonymized table below, we have replaced all the values in the 'Name' attribute and all the values in the 'Religion' attribute with a '*'.
2. **Generalization**: In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30', etc.

The next table shows the anonymized database.

Name	Age	Gender	County	Religion	Disease
*	20 < Age ≤ 30	Female	Surrey	*	Cancer
*	20 < Age ≤ 30	Female	Essex	*	Viral infection
*	20 < Age ≤ 30	Female	Hants	*	TB
*	20 < Age ≤ 30	Male	Sussex	*	No illness
*	20 < Age ≤ 30	Female	Hants	*	Heart-related
*	20 < Age ≤ 30	Male	Surrey	*	TB
*	Age ≤ 20	Male	Herts	*	Cancer
*	20 < Age ≤ 30	Male	Suffolk	*	Heart-related
*	Age ≤ 20	Male	Norfolk	*	Heart-related
*	Age ≤ 20	Male	Surrey	*	Viral infection

This data has 2-anonymity with respect to the attributes 'Age', 'Gender' and 'County' since for any combination of these attributes found in any row of the table there are always at least 2 rows with those exact attributes. The attributes available to an adversary are called "quasi-identifiers". Each "quasi-identifier" tuple occurs in at least k records for a dataset with k-anonymity.

Warnings

Because k-anonymization does not include any randomization, attackers can still make inferences about data sets that may harm individuals. For example, if the 19-year-old John from Surrey is known to be in the database above, then it can be reliably said that he has either cancer, a heart-related disease, or a viral infection.

k-anonymization is not a good method to anonymize high-dimensional datasets. k-anonymity can skew the results of a data set if it disproportionately suppresses and generalizes data points with unrepresentative characteristics.

The suppression and generalization algorithms used to k-anonymize datasets can be altered, however, so that they do not have such a skewing effect.

- **l-diversity: Privacy in data beyond k-anonymity**

It is a form of group-based anonymization that is used to preserve privacy in data sets by reducing the granularity of a data representation. This reduction is a tradeoff that results in some loss of effectiveness of data management or mining algorithms in order to gain some privacy. The l-diversity model is an extension of the k-anonymity model which reduces the granularity of data representation using techniques including generalization and suppression such that any given record maps onto at least k other records in the data. The l-diversity model handles some of the weaknesses in the k-anonymity model where protected identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group exhibit homogeneity. The l-diversity model adds the promotion of intra-group diversity for sensitive values in the anonymization mechanism.

Attacks on k-anonymity

While k-anonymity is a promising approach to take for group-based anonymization given its simplicity and wide array of algorithms that perform it, it is however susceptible to many attacks. When background knowledge is available to an attacker, such attacks become even more effective. Such attacks include:

- Homogeneity Attack: This attack leverages the case where all the values for a sensitive value within a set of k records are identical. In such cases, even though the data has been k-anonymized, the sensitive value for the set of k records may be exactly predicted.
- Background Knowledge Attack: This attack leverages an association between one or more quasi-identifier attributes with the sensitive attribute to reduce the set of possible values for the sensitive attribute. For example, Machanavajjhala, Kifer, Gehrke, and Venkatasubramanian (2007) showed that knowing that heart attacks occur at a reduced rate in Japanese patients could be used to narrow the range of values for a sensitive attribute of a patient's disease.

Formal Definition

Given the existence of such attacks where sensitive attributes may be inferred for k-anonymity data, the l-diversity method was created to further k-anonymity by additionally maintaining the diversity of sensitive fields.

The l-diversity Principle – An equivalence class is said to have l-diversity if there are at least l “well-represented” values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

Machanavajjhala et. al. (2007) define "well-represented" in three possible ways:

1. **Distinct l-diversity** – The simplest definition ensures that at least l distinct values for the sensitive field in each equivalence class.
2. **Entropy l-diversity** – The most complex definition defines Entropy of an equivalent class E to be the negation of summation of s across the domain of the sensitive attribute of p(E,s)log(p(E,s)) where p(E,s) is the fraction of records in E that have the sensitive value s. A table has entropy l-diversity when for every equivalent class E, Entropy (E) ≥ log (l).
3. **Recursive (c-l)-diversity** – A compromise definition that ensures the most common value does not appear too often while fewer common values are ensured to not appear too infrequently.

Aggarwal and Yu (2008) note that when there is more than one sensitive field the l-diversity problem becomes more difficult due to added dimensionalities.

t-closeness: A novel privacy-concept:

It is a further refinement of l-diversity group based anonymization that is used to preserve privacy in datasets by reducing the granularity of a data representation. This reduction is a tradeoff that results in some loss of effectiveness of data management or mining algorithms in order to gain some privacy. The t-closeness model extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

Attacks on l-diversity

This is useful because in real datasets attribute values may be skewed or semantically similar. However, accounting for value distributions may cause difficulty in creating feasible l-diverse representations. The l-diversity technique is useful in that it may hinder an attacker leveraging the global distribution of an attribute's data values in order to infer information about sensitive data values. Not every value may exhibit equal sensitivity, for example, a rare positive indicator for a disease may provide more information than a common negative indicator. Because of examples like this, l-diversity may be difficult and unnecessary to achieve when protecting against attribute disclosure. Alternatively, sensitive information leaks may occur because while l-diversity requirement ensures "diversity" of sensitive values in each group, it does not recognize that values may be semantically close, for example, an attacker could deduce a stomach disease applies to an individual if a sample containing the individual only listed three different stomach diseases

Formal definition

Given the existence of such attacks where sensitive attributes may be inferred based upon the distribution of values for l-diverse data, the t-closeness method was created to further l-diversity by additionally maintaining the distribution of sensitive fields. The original paper by Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian defines t-closeness as:

The t-closeness Principle: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

Charu Aggarwal and Philip S. Yu further state in their book on privacy-preserving data mining that with this definition, threshold t gives an upper bound on the difference between the distributions of the sensitive attribute values within an anonymized group as compared to the global distribution of values. They also state that for numeric attributes, using t-closeness anonymization is more effective than many other privacy-preserving data mining methods.

CONCLUSION: DATA DE-IDENTIFICATION: A GROWING NEED TO EMBED PRIVACY-ENHANCEMENT IN BIG DATA-USES.

- Deluge of data is a reality today due to their constant use in intellectual data-science-fields, medical science, technology & humanities enterprises.
- The instance of data being used by others without the possibility of individuals being identified can do wonders in this new data-driven age, by protecting the privacy of individuals, organizations & businesses and ensuring that the spatial location of minerals or endangered species is not publicly available.

REFERENCES:

1. Aggarwal, Charu C.; Yu, Philip S. (2008). "A General Survey of Privacy-Preserving Data Mining Models and Algorithms". Privacy-Preserving Data Mining – Models and Algorithms (PDF). Springer. pp. 11–52. ISBN 978-0-387-70991-8.
2. Li, Ninghui; Li, Tiancheng; Venkatasubramanian, S. (April 2007). "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007: 106–115. doi:10.1109/ICDE.2007.367856.
3. Machanavajjhala, Ashwin; Kifer, Daniel; Gehrke, Johannes; Venkatasubramaniam, Muthu Ramakrishnan (March 2007).
4. "Glossary of Statistical Terms: Quasi-identifier". OECD. November 10, 2005. Retrieved 29 September 2013.
5. Sweeney, Latanya. Simple demographics often identify people uniquely. Carnegie Mellon University, 2000. http://dataprivacylab.org/projects/identifiability/paper1.pdf
6. Anderson, Nate. Anonymized data really isn't—and here's why not. Ars Technica, 2009. http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/
7. Barth-Jones, Daniel C. The're-identification'of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (June 4, 2012) (2012).
8. Sweeney, Latanya, Akua Abu, and Julia Winn. "Identifying participants in the personal genome project by name." Available at SSRN 2257732 (2013).
9. Narayanan, Arvind and Shmatikov, Vitaly. Robust De-anonymization of Large Sparse Datasets.The University of Texas at Austin, 2008. https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
10. Rajeev Motwani and Ying Xu (2007). Efficient Algorithms for Masking and Finding Quasi-Identifiers (PDF). Proceedings of the Conference on Very Large Data Bases (VLDB).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Shabbir Anishbhai Bookseller
Quartesian
+91 – 895-146-4794
Shabbir.bookseller@quartesian.com