**Customer Identification Certification Report**

Adeline Casali

10/17/2023

DSE6003 Final Project

**Executive Summary**

Customer identification was masked, and the risk of identification was evaluated for the telecommunications customer survey dataset “Customer\_Data.xlsx”. The risk of customer re-identification was reviewed under four different scenarios using generally accepted probabilistic algorithms and scientific principles. It was determined that there is a tolerable risk of customers being identified in the shareable version of the dataset. Further recommendations to reduce the identification risk are provided in Appendix 3.

**Consequences of a Privacy Breach**

The consequences of a privacy breach are tremendously high, from both financial and reputational standpoints. Failure to adequately remove all personally identifying information and mask indirect identifiers to an acceptable risk of re-identification could result in an FCC violation or potential lawsuit (FCC, 2023). Companies have both legal and ethical requirements to prevent this from happening.

**Data Description**

The telecommunications customer survey dataset contained the information of 5000 customers. It included information on their geographic locations, income and spending habits, and various other pieces of sensitive personal information in conjunction with information on each customer’s phone usage. The variables chosen for the sharable version of the dataset were age, gender, marital status, years of education, and phone company tenure. This will allow our partner telecommunications company to analyze which combinations of age, gender, marital status, and education make the longest-term (and therefore “best”) customers. Once masked of personally identifiable information, each equivalence class contained between 2 and 209 customers, with a median of 44.

**Customer Risk Identification Methodology**

*Step 1*: We selected the following customer indirect variables or quasi-identifiers: gender, age, marital status, years of education, and phone company tenure. They were selected due to their utility to the partner telecommunication company for data analysis and marketing.

*Step 2*: Quality control of the data was performed to look for null or missing values, outliers, and other extreme records. None were found.

*Step 3*: Equivalence classes (EC) of patients were created as a group of records with the same combination of values for any given number of quasi-identifiers.

*Step 4*: An acceptable risk threshold was set at 0.33 (33%) because although the data does not contain health-related information, it does still contain sensitive personal information.

*Step 5*: The patient identification risk was measured under four possible attacks that can be made on the data by an intruder:

* *Scenario 1*: The intruder deliberately attempts to identify patient data
* *Scenario 2*: The intruder inadvertently identifies patient data
* *Scenario 3*: There is a data breach at the data recipient’s site and the data is exposed
* *Scenario 4*: An adversary launches a demonstration attack on the data; this occurs when an adversary wants to make a point of showing that a dataset can be re-identified

**Scenario 1: Deliberate Attack**

The main assumption in this scenario is the possibility that someone at the data recipient’s site will attempt to identify the data. For example, there may be a rogue staff member who wants to monetize the data for financial gain. The calculation of risk is based on the product of the following two probabilistic measures:

* Probability of attempt: Being conservative and assuming that the mitigating controls are low and the motives and capacity are high, we estimate the probability of attempt at 0.60 (60%) (El Emam, 2013).
* Probability of re-identification: The probability of correctly identifying a customer given an attack was computed directly from the dataset.

**Scenario 2: Inadvertent Data Attack**

The main assumption in this scenario is that someone at the data recipient’s site recognizes a patient in the dataset. For example, a sales representative working with the dataset may recognize an acquaintance, such as a relative or a neighbor, in the dataset through age and years of education. The calculation of risk is based on the product of the following two probabilistic measures:

* Probability of acquaintance: On average, people tend to have 150 friends (Dunbar, 1993). Additionally, it is estimated that 97% of people own a cell phone (Pew Research Center, 2021). Then, the probability of acquaintance is 1-(1-0.97)150= 100%.
* Probability of re-identification: The probability of correctly identifying a customer given that the adversary knows someone in the population covered by the dataset was computed directly from the dataset.

**Scenario 3: Data Breach**

This scenario can take place if the data recipient loses the dataset, or a data breach occurs. Based on current evidence, approximately 27% of healthcare providers covered under HIPAA have a reportable breach each year. We use this number to be conservative. The calculation of risk is based on the product of the following two probabilistic measures:

* Probability of a breach: 0.27 (27%)
* Probability of re-identification: The probability of correctly identifying a customer given a data breach has occurred was computed directly from the dataset.

**Scenario 4: Demonstration Attack**

This scenario can take place when the data is disclosed publicly. We assume that there is an intruder who has background information that can be used to launch an attack on the data, and that the intruder will attempt a customer identification attack. We are unaware of all the people who will be utilizing this dataset, so we must consider this as a plausible scenario. The calculation of this risk is determined by the probability of re-identification, which was computed directly from the dataset.

**Results**

*Diagnostics*

A table of numbers with numbers

Description automatically generated with medium confidence

*Figure 1: The percent of customers at each risk range for each scenario*

*Conclusions*

Risk is estimated to be tolerable for scenarios 1 and 3 (deliberate attack and data breach) but unacceptable for scenarios 2 and 4 (inadvertent and demonstration attack).

A red and white bar with black text

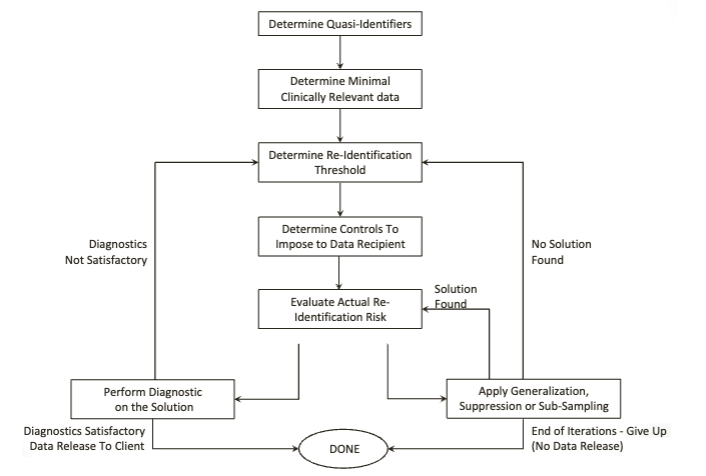
Description automatically generated

*Figure 2: Results from each scenario calculations*

**Appendix 1: Data Quality Control**

No outliers, null or missing values, or other extreme records were found within the dataset.

**Appendix 2: Risk Assessment Methodology**



*Figure 3: Recommended risk assessment methodology (El Emam et al., 2009)*

**Appendix 3: Recommendations**

We recommend that the ongoing delivery of data is based on secure procedures. Full documentation of the data recipient’s data storage and accessibility protocols must be assessed to further determine risks. Additionally, it may be useful to put into place a data-sharing agreement that prohibits re-identification attempts, an audit requirement, and the requirement to pass on these controls to any other party the data is shared with (Kniola, 2016). Finally, further grouping and masking could be done within the existing dataset, but it would greatly lower its utility.

**References:**

Dunbar, R. I. (1993). Coevolution of neocortical size, group size and language in humans. Behavioral and Brain Sciences, 16(4), 681–694. https://doi.org/10.1017/s0140525x00032325

Emam, K. E. (2013). Guide to the de-identification of Personal Health Information. CRC Press.

Emam, K. E., Dankar, F. K., Vaillancourt, R., Roffey, T., &amp; Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. The Canadian Journal of Hospital Pharmacy, 62(4). https://doi.org/10.4212/cjhp.v62i4.812

FCC. (2023). Protecting your personal data. Federal Communications Commission. https://www.fcc.gov/protecting-your-personal-data#:~:text=FCC%20rules%20protect%20customer%20proprietary,consumer%2C%20such%20as%20call%20waiting.

Kniola, L. (2016). (rep.). Calculating the Risk of Re-Identification of Patient-Level Data Using Quantitative Approach. Maidenhead, UK: PhUSE.

Pew Research Center. (2021, April 7). Mobile fact sheet. Pew Research Center: Internet, Science &amp; Tech. https://www.pewresearch.org/internet/fact-sheet/mobile/