*MEDINFO 2017: Precision Healthcare through Informatics*
*A.V. Gundlapalli et al. (Eds.)*

*doi:10.3233/978-1-61499-830-3-1303*

# Evaluation of the Anonymity and Utility of De-Identified Clinical Data Based on Japanese Anonymization Criteria

**Eizen Kimura[a], Satoshi Hasegawa[b], Koji Chida[b], Shoko Gamo[a], Satoshi Irino[c], Haku Ishida[d], Yukio Kurihara[e]**

[a] *Department of Medical Informatics, Ehime University Graduate School of Medicine, Touon, Ehime, Japan*
[b] *NTT Secure Platform Laboratories, Musashino, Tokyo, Japan*
[c] *Department of Nursing, Ehime Pref. University of Health Sciences, Tobe, Ehime, Japan*
[d] *Department of Medical Informatics & Decision Sciences, Yamaguchi University, Ube, Yamaguchi, Japan*
[e] *Division of Health Informatics, Medical School, Kochi University, Kyoto, Kyoto, Japan*

**Abstract**

*We analyze the deterioration of clinical data quality due to anonymization. The result shows that data quality remained high with micro-aggregation and also verify the availability of noise addition to prevent illegal re-identification by matching another personal data.*

***Keywords:***

Confidentiality, Anonymization, Privacy Preserving Data

## Introduction

The Amended Act on the Protection of Personal Informationin Japan, scheduled to go into full effect in the May of 2017, will open a way to provide medical records to third parties by de-identifying personal data. For de-identifying data, the law mandates the processing of personal information in accordance with standards stipulated by the rules of the Personal Information Protection Committee (PPC) [1] and the relevant guidelines. Currently, the PPC states that the five requirements must be met for the provision of data to a third party without personal agreement. At present, however, it is not clear what process will be sufficient to fulfill these rules because of the lack of technical guidelines for the anonymization process. The nature of an appropriate anonymization process may change depending on the content of the data and type of analysis. Our study examines use cases for clinical data to establish the best practice in anonymization.

## Methods

We collected 7,200 quasi-healthy individuals with age, sex, and laboratory items: Hb, CR, ALT, RBC, WBC, aspartate aminotransferase, total protein, blood urea nitrogen, total cholesterol, and gamma glutamyl transpeptidase (GGT). Using the data, we examine various statistics, mean, standard deviation, quartile, the Mann–Whitney U-test (MU) and the Kolmogorov–Smirnov (KS) test for each group of the same age class and sex as a case study. Based on the age hierarchy required by researchers, ages were converted into four classes using 5-year intervals: 50–54, 55–59, 60–64, and 65–69 years.

As a requirement of the five rules, we evaluate k-anonymity assuming age and sex are quasi-identifiers. We also assumed that the exceptional description of a single attribute falls within three standard deviations (SDs) of all records and is within top and bottom 2.5% data within each same-sex group 5-year interval in age. We applied Maximum Distance to Average Vector (MDAV[2]) for micro-aggregation. To reduce the possibility of identification by collating medical records with other data, we used Pk-anonymization[3] with noise addition and the Post Randomization Method (PRAM[4]).

## Results

All data satisfied k=10 in the age and sex pairs. Therefore, k-anonymization processing was unnecessary for age and sex. We confirmed the errors were exceedingly small even if micro-aggregation withing  3 SDs is applied. The errors of RBC are smaller than those of GGT, the de-identified data with noise-addition cannot be used in either the basic statistics or hypothesis tests under existing conditions. The lower error rate of RBC compared with GGT may be attributable the wide range of GGT values, compared to the RBC values. Also, RBC has a closer distribution to the normal distribution.

## Conclusion

We instantiated de-identified clinical data considering the requirements of Japanese anonymity criteria and verified the availability. Anonymization using only micro-aggregation maintained good data quality. However, we found that many sample data are required to obtain the statistical result with high quality for de-identified data with noise addition based on Pk-anonymity. In a future study, we will re-evaluate our results after improving the noise-addition algorithm and increasing the target data size, assuming large data processing.

## References

[1] Personal Information Protection Commision JAPAN, *https://www.ppc.go.jp/en/legal/* (2017).

[2] J. Domingo-Ferrer and V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery* **11** (2005), 195-212.

[3] D. Ikarashi, R. Kikuchi, K. Chida, and K. Takahashi, k-Anonymous Microdata Release via Post Randomisation Method, in: *International Workshop on Security*, Springer, 2015, pp. 225-241.

[4] P. Kooiman, L.C.R.J. Willenborg, and J.M. Gouweleeuw, *PRAM: a Method for Disclosure Limitation of Microdata*, CBS, 1997.

**Address for correspondence**

Eizen Kimura, ekimura@m.ehime-u.ac.jp

Department of Medical Informatics, Ehime University Graduate School of Medicine