BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Exploring the tradeoff between data privacy and utility with a clinical data analysis use case

Eunyoung Im[1,2], Hyeoneui Kim[1,2,3*], Hyungbok Lee[1,5], Xiaoqian Jiang[4] and Ju Han Kim[5,6]

Abstract

Background Securing adequate data privacy is critical for the productive utilization of data. De-identification, involving masking or replacing specific values in a dataset, could damage the dataset's utility. However, finding a reasonable balance between data privacy and utility is not straightforward. Nonetheless, few studies investigated how data de-identification efforts affect data analysis results. This study aimed to demonstrate the effect of different de-identification methods on a dataset's utility with a clinical analytic use case and assess the feasibility of finding a workable tradeoff between data privacy and utility.

Methods Predictive modeling of emergency department length of stay was used as a data analysis use case. A logistic regression model was developed with 1155 patient cases extracted from a clinical data warehouse of an academic medical center located in Seoul, South Korea. Nineteen de-identified datasets were generated based on various de-identification configurations using ARX, an open-source software for anonymizing sensitive personal data. The variable distributions and prediction results were compared between the de-identified datasets and the original dataset. We examined the association between data privacy and utility to determine whether it is feasible to identify a viable tradeoff between the two.

Results All 19 de-identification scenarios significantly decreased re-identification risk. Nevertheless, the de-identification processes resulted in record suppression and complete masking of variables used as predictors, thereby compromising dataset utility. A significant correlation was observed only between the re-identification reduction rates and the ARX utility scores.

Conclusions As the importance of health data analysis increases, so does the need for effective privacy protection methods. While existing guidelines provide a basis for de-identifying datasets, achieving a balance between high privacy and utility is a complex task that requires understanding the data's intended use and involving input from data users. This approach could help find a suitable compromise between data privacy and utility.

Keywords Data privacy, Data utility, Data de-identification, Clinical data analysis, ARX tool

*Correspondence:
Hyeoneui Kim
ifilgood@snu.ac.kr
[1]College of Nursing, Seoul National University, Seoul, South Korea
[2]Center for World-leading Human-care Nurse Leaders for the Future by Brain Korea 21 (BK 21) four project, College of Nursing, Seoul National University, Seoul, South Korea
[3]The Research Institute of Nursing Science, Seoul National University, Seoul, South Korea
[4]School of Biomedical Informatics, UTHealth, Houston, TX, USA
[5]Seoul National University Hospital, Seoul, South Korea
[6]College of Medicine, Seoul National University, Seoul, South Korea

Background

Clinical data gathered through Electronic Health Records (EHR) is an invaluable asset for producing meaningful insights into patient care and healthcare service management. However, as this data includes sensitive personal information, there is a heightened risk of financial or social damage to individuals if their health data is improperly disclosed [1, 2]. To address these concerns, many countries have implemented stringent regulations to safeguard patient privacy while still enabling the efficient use of data for health advancements [3]. In the United States, for example, the Health Insurance Portability and Accountability Act (HIPAA) sets forth provisions for data protection and usage [4]. Similarly, the General Data Protection Regulation (GDPR) offers a comprehensive data privacy framework within the European Union [5]. Additionally, South Korea's Personal Information Protection Act delineates the guidelines for secure and permissible data handling [6].

The growing imperative for data privacy has spurred significant progress in privacy-preserving technologies. Differential Privacy (DP) safeguards data by integrating controlled random noise, thus ensuring individual data points remain confidential while aggregate analysis remains accurate [7]. In the biomedical field, DP is extensively employed in data query systems; the noise integrated into query responses helps protect sensitive inquiries pertaining to uncommon cases [8, 9]. Current research in DP focuses on solving complex problems such as determining optimal privacy budgets and noise levels to balance confidentiality with data utility [8, 10, 11].

Homomorphic Encryption (HE) represents a breakthrough in cryptography for preserving privacy, enabling computations on encrypted data without altering the original values [12]. Recent research has validated the practicality of performing data analysis using HE [13–15]. Nonetheless, HE has not become mainstream in healthcare applications, primarily due to its substantial computational demands, intricate implementation, and the limited range of analytics that can be performed on data in its encrypted form [12, 16].

Blockchain technology, recognized for its immutable, decentralized, and transparent nature [17], is gaining attention as an innovative approach for data privacy [18–20]. Despite this interest, the real-world application of blockchain is contingent upon enhancements in its capacity to process substantial data volumes, simplification of its implementation, and resolution of related regulatory challenges [21–24].

When preparing datasets with personal health information for secondary analysis, the prevailing practice is to mitigate the risk of re-identification of the subjects in the dataset by employing stringent de-identification procedures [25, 26]. This involves the removal of direct identifiers that can uniquely pinpoint individual subjects within the dataset and altering quasi-identifiers, which alone do not identify subjects but could do so when merged with other data sources. Furthermore, the process considers sensitive information that, despite not directly identifying subjects, could have detrimental effects if disclosed, ensuring such data is also considered during the de-identification process.

The leading method for data de-identification employs strategies like K-anonymity, L-diversity, and T-closeness to modify data. K-anonymity safeguards against linkage attacks by ensuring that there are at least K identical records for any set of quasi-identifiers within a dataset, making it impossible to distinguish one individual from K-1 others [27]. In line with this, South Korea's data publishing guidelines recommend adhering to a minimum of 'K=3' for K-anonymity [28, 29]. Additionally, L-diversity mandates a sensitive variable must have at least L distinct values, thereby offering protection against homogeneity attacks [30]. T-closeness, on the other hand, ensures that the distribution of a sensitive variable within any subset of the dataset closely approximates the distribution of that variable of the entire dataset, adhering to a specified threshold [31]. T-closeness prevents the likelihood that knowledge of the variable's distribution could be exploited to reveal an individual's identity [31]. The process of de-identification, which often involves masking or altering certain data values, can result in information loss and potentially reduce the utility of the dataset [32].

Determining the optimal threshold between data privacy and utility remains a complex challenge. Several studies have investigated how various de-identification strategies, specifically K-anonymity, L-diversity, and T-closeness, influence data utility. This is typically assessed by comparing the analytical results of de-identified datasets with those derived from the original dataset. Some researchers advocate that the privacy enhancements are overshadowed by a substantial reduction in data utility [33, 34], while others argue that such utility loss might not be as severe as some studies imply [35]. However, these studies evaluated each de-identification technique in isolation, often resorting to simplified models that fail to fully capture the complexities of real-world data use, and led to mixed conclusions [34, 35].

Moreover, the insights offered by such research into the tangible effects of data de-identification on actual data analysis tasks are somewhat restricted. This is because the analyses were either performed using overly simplistic examples [28, 34] or on public datasets that have already undergone some form of de-identification [35, 36], or focusing on theoretical aspects [37]. Therefore, there is a need for more intricate research that closely mirrors the complexities of real-life data analytics tasks

Table 1 The variables extracted from the clinical data warehouse

Variables	Description
Sex	Sex of the patients
Age	Age of the patients in years
Acuity level	Patient's acuity level classified based on Korean Triage and Acuity Scale (KTAS) level
Number of consults	Number of consults requested from the emergency department
Inter-hospital communication	Prior communication between medical staff at the time of transfer
Sending hospital	Name and address of the healthcare facility that the patient was transferred from
Primary diagnosis	Patients' main diagnosis coded with International Classification of Diseases (ICD 10)
Treatment outcome	Follow-up measures according to patients' progress after admission to the emergency department (e.g., Discharge, Admission, Transfer)
Length of stay	Duration that the patient stayed at ED, measured in hours

and considers the multifaceted nature of data utility and privacy in actual applications.

This study explores the effects of different de-identification strategies on clinical datasets prepared for secondary analysis, with a focus on their implications for practical data analysis tasks. The aims of this study are twofold: firstly, to assess the effects of de-identification on both the dataset's integrity and the outcomes of data analyses; and secondly, to ascertain if discernible trends emerge from the application of various de-identification techniques that could guide the establishment of a feasible balance between data privacy and data utility.

Methods

Data analysis use case

This study explores the impact of various de-identification techniques on datasets and their subsequent analysis results using a data analytic use case. The analytic use case involved predicting the Length of Stay (LOS) of high-acuity patients transferred to the emergency department (ED) of an academic medical center located in Seoul, South Korea. LOS in the ED serves as a crucial quality metric for ED services [38–40]. In Korea, an ED LOS under six hours is considered optimal [41]. Nonetheless, the overcrowding issues prevalent in tertiary hospital EDs elevate the risk of prolonged ED stays for patients transferred from other facilities for specialized care [42, 43]. Understanding the factors affecting the ED LOS of transferred high-acuity patients is essential to providing timely care. The authors, HK and HL, previously developed a model to predict ED LOS using logistic regression, Random Forest, and Naïve Bayes techniques [44]. Building on insights from this earlier research, the current use case was crafted to develop a logistic regression model to predict ED LOS based on variables including the patient's sex, age, medical conditions, the type and location of the transferring hospital, and the treatment outcomes.

Table 2 The fraction of the records potentially identifiable by single or combinations of variables

Single or combinations of variables	% Records potentially identifiable
Sex	0%
Age	0.52%
Primary diagnosis	17.75%
Sex + Primary diagnosis	25.02%
Sending hospital	27.71%
Sex + Sending hospital	38.10%
Sex + Age + Primary diagnosis	83.72%
Age + Sending hospital	88.40%
Primary diagnosis + Sending hospital	93.59%
Age + Sending hospital + Primary diagnosis	98.70%
Sex + Age + Sending hospital + Primary diagnosis	98.70%

Dataset

The prediction model for ED LOS was developed using data from 1,155 patients who were transferred to the study site's ED between January 2019 and December 2019. Patient demographics, clinical details, and transfer-related information were extracted from the study site's Clinical Data Warehouse (CDW). The variables collected for this study are listed in Table 1.

De-identification of the datasets

Developing de-identification scenarios

Identifiers such as patient names and medical record numbers were removed. Quasi-identifiers play a critical role in de-identification as they form the foundation for assessing the adequacy of de-identification efforts and undergo most data transformations. To select the variables to test as quasi-identifiers, we first examined the extent to which each variable could uniquely link to individual subjects within the dataset, potentially identifying them. Table 2 displays the percentage of subjects in the dataset uniquely linked to either a single variable or a combination of variables. For instance, the *sending hospital* and *primary diagnosis* were uniquely linked to 27.71% and 17.75% of the subjects, respectively, and their combination linked up to 94% of the subjects. Consequently, information regarding the *sending hospital* and the *primary diagnosis*, coded using the International Classification of Disease (ICD) [45], were utilized as quasi-identifiers, along with *sex* and *age*, which are commonly considered quasi-identifiers in various de-identification efforts [4, 46]. Treatment outcomes were identified as sensitive information. We developed 19 de-identification scenarios by varying the quasi-identifiers and sensitive information, and applying diverse configurations

of privacy-preserving techniques such as K-anonymity, L-diversity, and T-closeness to each scenario.

Data transformation for de-identification

De-identification was performed using ARX, a publicly accessible and well-validated data anonymization tool that supports various de-identification methods [47–49]. We employed generalization and micro-aggregation techniques to modify the quasi-identifiers, both aimed at reducing the risk of re-identification by transforming original data into more general values. Generalization involves building a hierarchy for the given values by specifying minimum and maximum generalization levels. Generalization involves creating a hierarchy of values by specifying minimum and maximum levels, which can be adjusted based on criteria such as the number of digits masked in zip codes, size of intervals for *age*, condensation of 5-point Likert scores to 3-point scales, and generalization of full dates to broader time units such as week, month, or year [50]. Micro-aggregation, on the other hand, assigns representative values for alphanumeric data, such as using the mode for *sex* and the mean for *age* [50].

In our de-identification process, quasi-identifiers such as the *sending hospital* and *primary diagnosis* were transformed using generalization, while *sex* was modified through micro-aggregation. *Age* was subjected to both generalization and micro-aggregation. The generalization hierarchy for *age* included three levels with intervals of 5, 10, and 30 years respectively. For micro-aggregation, mean *age* values were used. The *primary diagnosis* was generalized into two levels based on higher-level ICD codes. For instance, a *primary diagnosis* with the ICD code I20.0, representing *unstable angina*, was generalized to I20 (i.e., *angina pectoris*) at level 1, and further to I20-I25 (i.e., *ischemic heart diseases*) at level 2. Generalization of the *sending hospital* also included two levels, where a specific facility such as "Hanmaeum Clinic in Jongno-gu, Seoul city" was generalized to the county level as "facility in Jongno-gu" at level 1 and then to the city level as "facility in Seoul" at level 2. For *sex*, micro-aggregation was employed, setting the mode as the representative value.

K-anonymity, L-diversity, and T-closeness were employed concurrently with specific parameters set for each: K and L were both set at 3, and T was set at 0.5. K-anonymity was specifically applied to quasi-identifiers to ensure that each individual is indistinguishable from at least two others. L-diversity and T-closeness, on the other hand, were applied to the variable designated as sensitive, ensuring that sensitive information is both sufficiently diverse and closely aligned with the overall distribution of the dataset. Table 3 details these 19 de-identification scenarios.

Data transformation was carried out in ARX according to the de-identification scenarios outlined in Table 3. ARX provides options to adjust additional transformation parameters: the *suppression limit*, which sets the maximum proportion of records that can be omitted from the original dataset; *approximation*, which prioritizes solutions with shorter execution times; and *precomputation*, which determines the threshold for the fraction of unique data values in the dataset [50]. For this study, we utilized the default settings in ARX, where the *suppression limit* was set to 100%, and both *approximation* and *precomputation* features were disabled.

During execution, ARX evaluated various combinations of generalization and micro-aggregation levels to meet the requirements for K-anonymity, L-diversity, and T-closeness, ultimately recommending an optimal solution based on the balance between minimizing re-identification risk and preserving data utility. Figure 1 displays a screenshot of the data transformation solutions for the scenario where *age*, *primary diagnosis*, and *sending hospital* were designated as quasi-identifiers. Ultimately, we produced 19 versions of de-identified datasets, each based on the transformation solution that ARX identified as optimal.

Examination of the de-identified datasets

We reviewed the reduction in re-identification risk and the data utility scores that ARX estimated for the 19 de-identified datasets. To assess the similarity between each de-identified dataset and the original dataset, we employed Earth Mover's Distance (EMD) [51]. Additionally, we calculated the dataset retention ratio. This metric is derived by dividing the number of data points in the transformed dataset by the number of data points in the original dataset. EMD and dataset retention ratio quantitatively evaluate the dissimilarity between the original dataset and the de-identified datasets, offering insights into how much the data has been altered through de-identification.

Testing the effects of de-identification on ED LOS prediction

Variable creation for predictive modeling

To construct a logistic regression model for predicting ED LOS, we defined outcome and predictor variables. ED LOS, the outcome variable, was dichotomized into two categories: 6 h or less, and more than 6 h. We identified 13 predictors, including patient sex, age, medical conditions, treatment outcome, and the sending hospital type. *Age*, *sending hospital location*, and *treatment outcome* were dichotomized. Five dummy variables were created from *primary diagnosis* to represent *high priority disease*, *neoplastic disease*, *circulatory disease*, *respiratory disease*, and *injury-related visits*. The *sending hospital type*

Table 3 Data de-identification scenarios

De-identification scenario	Sex	Age	Primary diagnosis	Sending hospital	Treatment outcome
1	Micro-aggregation (mode)	Micro-aggregation (mean)	Generalization	Generalization	L-diversity
2	Micro-aggregation (mode)	Generalization	Generalization	Generalization	L-diversity
3	Micro-aggregation (mode)	Micro-aggregation (mean)	Generalization	Generalization	L-diversity, T-closeness
4	Micro-aggregation (mode)	Generalization	Generalization	Generalization	L-diversity, T-closeness
5	Micro-aggregation (mode)	Micro-aggregation (mean)	Generalization	L-diversity	L-diversity
6	Micro-aggregation (mode)	Generalization	Generalization	L-diversity	L-diversity
7	Micro-aggregation (mode)	Micro-aggregation (mean)	Generalization	L-diversity, T-closeness	L-diversity, T-closeness
8	Micro-aggregation (mode)	Generalization	Generalization	L-diversity, T-closeness	L-diversity, T-closeness
9	Micro-aggregation (mode)	Micro-aggregation (mean)	L-diversity	Generalization	L-diversity
10	Micro-aggregation (mode)	Generalization	L-diversity	Generalization	L-diversity
11	Micro-aggregation (mode)	Micro-aggregation (mean)	L-diversity, T-closeness	Generalization	L-diversity, T-closeness
12	Micro-aggregation (mode)	Generalization	L-diversity, T-closeness	Generalization	L-diversity, T-closeness
13	Micro-aggregation (mode)	Generalization	L-diversity	L-diversity	L-diversity
14	Micro-aggregation (mode)	Generalization	L-diversity, T-closeness	L-diversity, T-closeness	L-diversity, T-closeness
15	Micro-aggregation (mode)	Generalization	-	-	L-diversity
16	Micro-aggregation (mode)	Generalization	-	-	L-diversity, T-closeness
17	Micro-aggregation (mode)	-	Generalization	Generalization	L-diversity, T-closeness
18	-	Generalization	Generalization	Generalization	L-diversity, T-closeness
19	-	-	Generalization	Generalization	L-diversity, T-closeness

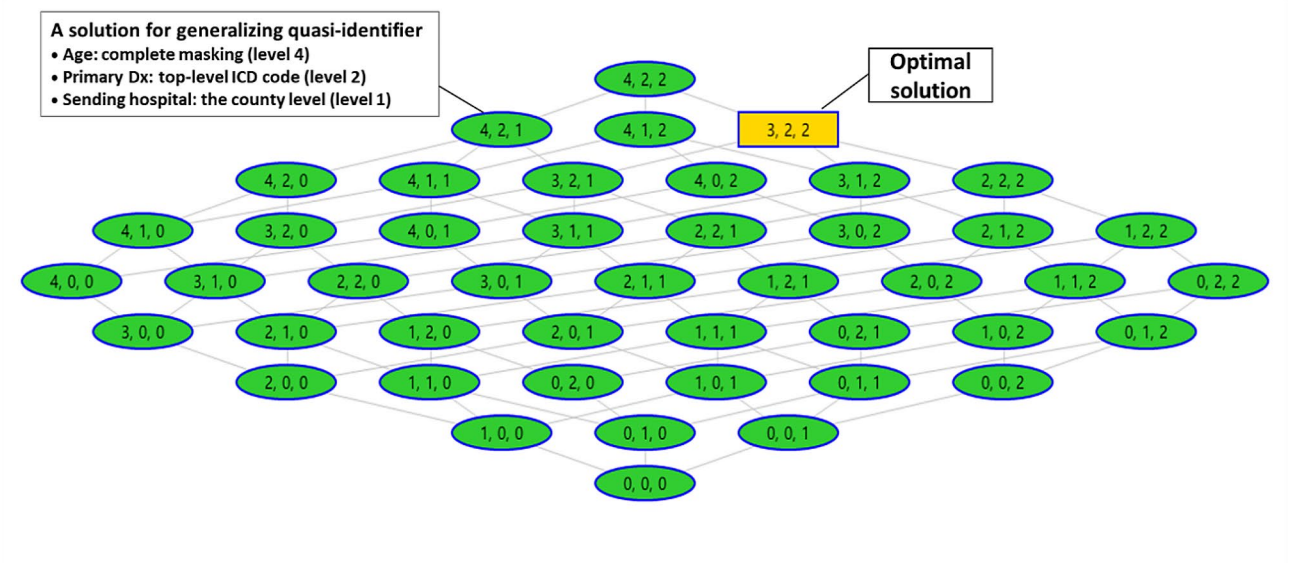


Fig. 1 The data transformation solutions suggested by ARX

Table 4 The definitions of the variables used in the logistic regression analyses

Variable names	Descriptions (encoding)
Sex	Sex of the patients (0 = Male, 1 = Female)
Age	Age of the patients in years (0 = < 60yrs, 1 = ≥ 60yrs)
Acuity level	Severity classification according to Korean Triage and Acuity Scale (KTAS) level (0 = Level 1, 1 = Others)
Number of consults	The number of consults (0 = < 3, 1 = ≥ 3)
Inter-hospital communication	Prior communication between medical staff at the time of transfer (0 = No, 1 = Yes)
Sending hospital location	Location of the sending hospital (0 = Inside seoul, 1 = Outside seoul)
High priority disease	Classification of high priority diseases included in the Korean High Priority Diseases Classification Standards (0 = No, 1 = Yes)
Neoplastic disease	Presence of neoplasm diseases (0 = No, 1 = Yes)
Circulatory disease	Presence of circulatory diseases (0 = No, 1 = Yes)
Respiratory disease	Presence of respiratory diseases (0 = No, 1 = Yes)
Sending hospital type	Type of the sending hospital according to patient's length of stay (0 = Short-term care facility, 1 = Long-term care facility)
Injury-related visits	Reason/types for visiting the emergency room due to illness or injury (0 = Disease, 1 = Injury)
Treatment outcome	Whether additional plans were established for patient treatment results (0 = Discharge/ Against medical advice (AMA) discharge, 1 = Admission/ Procedure/Operation/Transfer to other hospitals)
ED LOS	The length of stay at Emergency Department (0: LOS <= 6 h, 1: LOS > 6 h)

was derived from the *sending hospital information*. These variables, detailed in Table 4, were consistently defined across all 19 de-identified datasets as well as the original dataset to facilitate comparative analyses.

Data analysis

After defining the outcome and predictor variables for logistic regression, we examined their distributions across the 19 de-identified datasets and the original dataset. To assess the differences in variable distributions, we utilized the proportion test [52]. Subsequently, logistic regression analysis was conducted using both the de-identified and original dataset. The predictive performance of these models was evaluated using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. We compared the AUC scores (AUROC) of the logistic regression models derived from the 19 de-identified datasets to that from the original dataset, employing the DeLong test [53]. Additionally, we analyzed the differences in the odds ratios of the predictors and their statistical significance to assess any impact the de-identification process might have had on the predictive capability of the models. All analyses were performed using R (version 4.0.4) [54].

Results

Data transformation configurations applied for the de-identification of the datasets

Table 5 displays the optimal configurations for data transformation used in the 19 de-identified datasets. Variables subjected to generalization or micro-aggregation were

Table 5 The data transformation configurations applied to the de-identified datasets

De-identified dataset numbers	Sex	Age	Primary diagnosis	Sending hospital	Treatment outcome
1	Micro -mode	Micro -mean	Gen -level 2	Gen -level 2	SI
2	Micro -mode	Gen -level 3	Gen -level 2	Gen -level 2	SI
3	Micro -mode	Micro -mean	Gen -level 2	Gen -level 2	SI
4	Micro -mode	Gen -level 3	Gen -level 2	Gen -level 2	SI
5	Micro -mode	Micro -mean	Gen -level 2	SI	SI
6	Micro -mode	Gen -level 3	Gen -level 2	SI	SI
7	Micro -mode	Micro -all masking	Gen -all masking	SI	SI
8	Micro -mode	Gen -all making	Gen -all masking	SI	SI
9	Micro -mode	Micro -mean	SI	Gen -level 1	SI
10	Micro -mode	Gen -level 3	SI	Gen -level 1	SI
11	Micro -mode	Micro -mean	SI	Gen -level 2	SI
12	Micro -mode	Gen -level 3	SI	Gen -level 2	SI
13	Micro -mode	Gen -level 1	SI	SI	SI
14	Micro -mode	Gen -level 3	SI	SI	SI
15	Micro -mode	Gen -level 1			SI
16	Micro -mode	Gen -level 1			SI
17	Micro -mode		Gen -level 2	Gen -level 2	SI
18		Gen -level 3	Gen -level 2	Gen -level 2	SI
19			Gen -level 2	Gen -level 2	SI

Note. Micro: micro-aggregation, Gen: generalization, SI: sensitive information

designated as quasi-identifiers. Sensitive information is identified as 'SI' within the table. It is important to note that empty cells signify that the corresponding variable was treated as non-sensitive information in the specific dataset.

The de-identified datasets

Table 6 displays the re-identification reduction rates, ARX utility scores, EMD scores, and dataset retention ratios for the 19 transformed datasets. Additionally, the table presents the number of records retained post-transformation and the number of predictor variables generated. The ARX utility score reflects the extent of information loss, with a higher score indicating lower utility. It is important to note that the baseline re-identification risk varied among the datasets due to differences in the configuration of quasi-identifiers.

Overall, all 19 de-identification scenarios significantly reduced re-identification risk. However, the data transformation processes involved in de-identification led to record suppression and complete masking of variables used as predictors, thereby compromising dataset utility. Notably, except for three datasets (13, 15, 16), which used only *sex* and *age* as quasi-identifiers, there was a loss of one or more predictor variables. Datasets 13, 15, and 16 demonstrated the highest retention ratios and the lowest ARX utility and EMD scores, indicating minimal information loss and the highest similarity to the original dataset, thus reflecting superior dataset utility. They also exhibited the lowest baseline and post-transformation re-identification risks.

Datasets 7 and 8 underwent a transformation under the most complex de-identification scenarios, employing three quasi-identifiers and applying both L-diversity and T-closeness to two sensitive variables. Although these datasets achieved complete re-identification risk reduction, the extensive data transformation allowed only seven predictor variables to be generated. The de-identification scenarios 1 and 3, 2 and 4, and 13, 15, and 16 shared identical configurations of quasi-identifiers but varied in the L-diversity and T-closeness conditions applied to sensitive information, resulting in identical de-identified datasets (see Table 3).

Table 7 details the differences in variable distribution between each transformed dataset and the original dataset. As expected, variables designated as quasi-identifiers underwent the most transformation, leading to significant changes. Variables derived from these quasi-identifiers, such as *sending hospital type, circulatory disease*, and *high priority disease*, also exhibited notable distributional changes.

The prediction results

Logistic regression models were developed using both the original dataset and 19 de-identified datasets. The complete masking of variables classified as quasi-identifiers in some de-identified datasets resulted in differences in the number and types of predictors available for

Table 6 The features of the de-identified datasets

Dataset numbers	Re-identification risk -before	Re-identification risk -after	Re-identification risk reduction rate	ARX utility score	EMD	# of records retained for logistic regression	# of predictors retained for logistic regression	Dataset retention ratio
1	0.993	0.064	0.936	0.722	62.346	547	11	0.401
2	0.993	0.076	0.924	0.807	62.559	396	11	0.290
3	0.993	0.064	0.936	0.722	62.346	547	11	0.401
4	0.993	0.076	0.924	0.807	62.559	396	11	0.290
5	0.908	0.044	0.952	0.485	61.746	954	12	0.762
6	0.908	0.059	0.935	0.599	62.017	765	12	0.611
7	0.908	0.000	1.000	1.000	61.118	1119	7	0.522
8	0.908	0.000	1.000	1.000	61.118	1119	7	0.522
9	0.963	0.059	0.939	0.500	61.623	910	12	0.727
10	0.963	0.085	0.911	0.600	61.945	756	12	0.604
11	0.963	0.002	0.998	0.890	62.542	1155	9	0.692
12	0.963	0.002	0.998	0.846	62.737	1155	9	0.692
13	0.135	0.014	0.897	0.449	61.414	1113	13	0.964
14	0.135	0.002	0.986	0.654	61.521	1052	12	0.841
15	0.135	0.014	0.897	0.449	61.414	1113	13	0.964
16	0.135	0.014	0.897	0.449	61.414	1113	13	0.964
17	0.965	0.064	0.934	0.749	63.512	547	11	0.401
18	0.991	0.076	0.924	0.749	62.558	396	11	0.290
19	0.943	0.064	0.932	0.639	63.498	547	11	0.401

Note. The number of records in the original dataset: 1155, the number of predictors for logistic regression: 13

Table 7 The difference in the variable distributions between the de-identified datasets and the original dataset

Dataset numbers	Male	Age < 60	Acuity level	Number of consults	Inter hospital communication	Sending hospital location	High priority disease	Neo plastic disease	Circulatory disease	Respiratory disease	Sending hospital type	Injury-related visits	Treatment outcome: admission	LOS > 6 (outcome variable)
1	***	***			**	***	NA		**		NA	*		
2		***			**	***	NA		***		NA	**		
3	***	***			**	***	NA		**		NA	*		
4		***			**	***	NA		***		NA	**		
5	***	**					NA							
6		**					NA		*					
7	NA	NA					NA	NA	NA	NA				
8	NA	NA					NA	NA	NA	NA				
9	*	***				***					NA			
10		**				***					NA			
11	NA	NA				NA					NA			
12	NA	NA				NA					NA			
13	***													
14	NA													
15	***													
16	***													
17	***				**	***	NA		**		NA	*		
18		***			**	***	NA		***		NA	**		
19					**	***	NA		**		NA	*		

*$p<0.05$, **$p<0.01$, ***$p<0.001$, NA: not comparable because the variable was lost in the de-identified dataset.

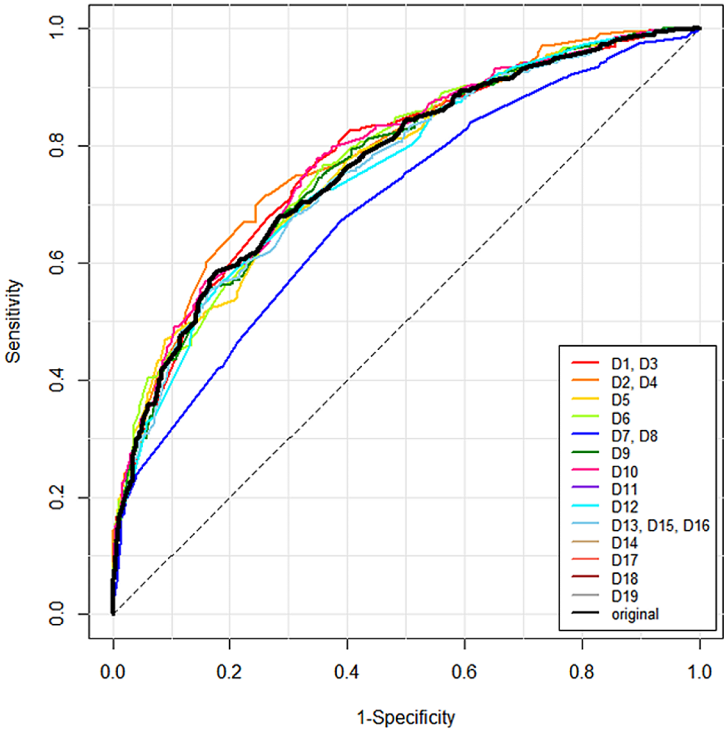
constructing the logistic regression models. Additionally, the number of records included in the regression analysis varied due to record suppression associated with the de-identification process. Figure 2 illustrates the ROC curves and the AUC values for all 20 datasets. The AUC values ranged from 0.695 to 0.787. The models generated from datasets 7 and 8, which only retained seven predictors due to extensive data masking, exhibited a statistically significant difference in AUC when compared to the original dataset, with a p-value of 0.002. For the models derived from the other datasets, no significant differences in AUC values were observed.

Figure 3 displays the Odds Ratios (OR) for predictors from selected datasets. Datasets 13, 15, and 16 were chosen because they retained all 13 predictor variables (Fig. 3(a)). Dataset 9 was selected for having the next highest number of predictors (N=12) and for utilizing three quasi-identifiers: the *sending hospital*, which is identified as the most revealing variable in Table 2, along with *sex* and *age*, which are commonly used as quasi-identifiers (Fig. 3(b)). Dataset 19 was also included because it was configured using only the *sending hospital* and *primary diagnosis* as quasi-identifiers (Fig. 3(c)). The ORs for all 19 datasets are detailed in Additional file 1: Figure S1.

As depicted in Fig. 3(a), the original dataset and de-identified datasets 13, 15, and 16 showed comparable prediction outcomes, with *sex* being the only predictor that displayed an OR notably different from the original dataset; however, it was not statistically significant in either model. Figure 3(b) indicates that the ORs of the 12 predictors in dataset 9 were similar to those in the original dataset, although the OR for injury-related visits became insignificant. In contrast, dataset 19, which excluded two predictors, showed more pronounced differences in the ORs of the 11 remaining predictors (Fig. 3(c)). Additionally, *neoplastic disease* and *respiratory disease*, significant predictors in the original dataset, became insignificant in dataset 9, while *injury-related visits*, previously insignificant, became significant (Fig. 3(c)).

Data utility vs. data privacy

Figure 4 presents the correlations between re-identification risk reduction rates, ARX utility scores, EMD, and dataset retention ratios. There is a significant correlation between the re-identification reduction rate and the ARX utility score, indicating that greater reductions in re-identification risk are typically accompanied by larger losses of information. Conversely, the re-identification reduction rate exhibits a slight negative correlation with both EMD and dataset retention ratio; however, these correlations are not statistically significant.



Dataset number	AUC	# of records retained for logistic regression	# of predictors retained for logistic Regression
1, 3	0.777	547	11
2, 4	0.787	396	12
5	0.767	954	12
6	0.774	765	12
7, 8	0.695*	1119	7
9	0.768	910	12
10	0.780	756	12
11	0.758	1155	9
12	0.758	1155	9
13, 15, 16	0.759	1113	13
14	0.767	1052	12
17	0.772	547	11
18	0.780	396	11
19	0.772	547	11
Original	**0.761**	**1155**	**13**

** Differ significantly from the original AUC at a 95% significance level when tested with the DeLong test*

Fig. 2 The number of records and predictors included in each model and the model performance

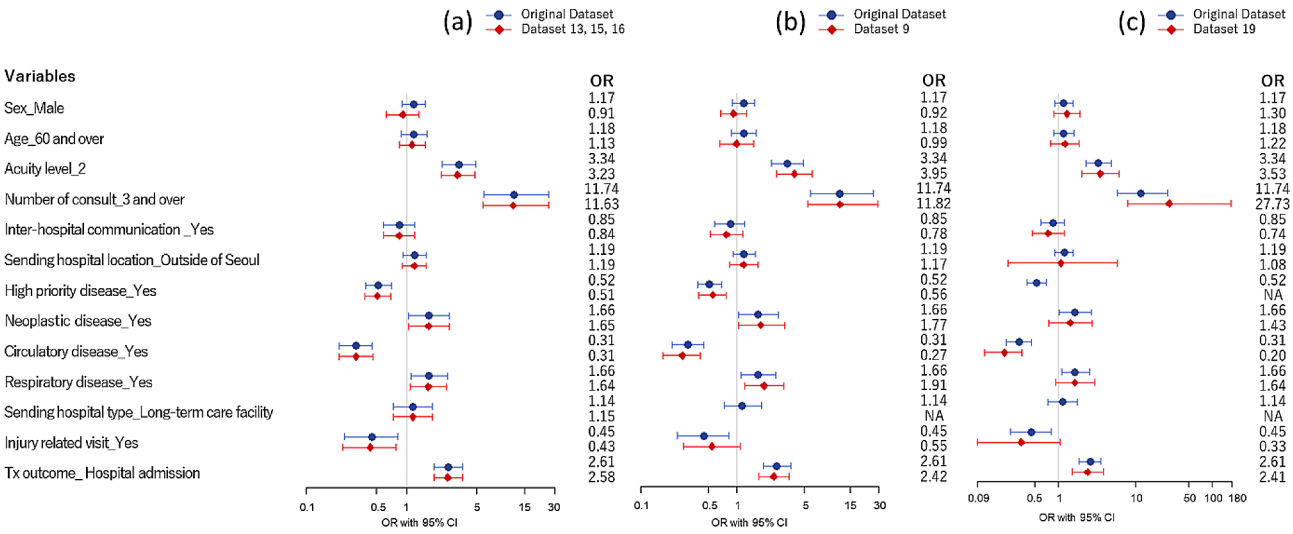


Fig. 3 The Odds-Ratios of the predictors from the original dataset and the selected de-identified datasets

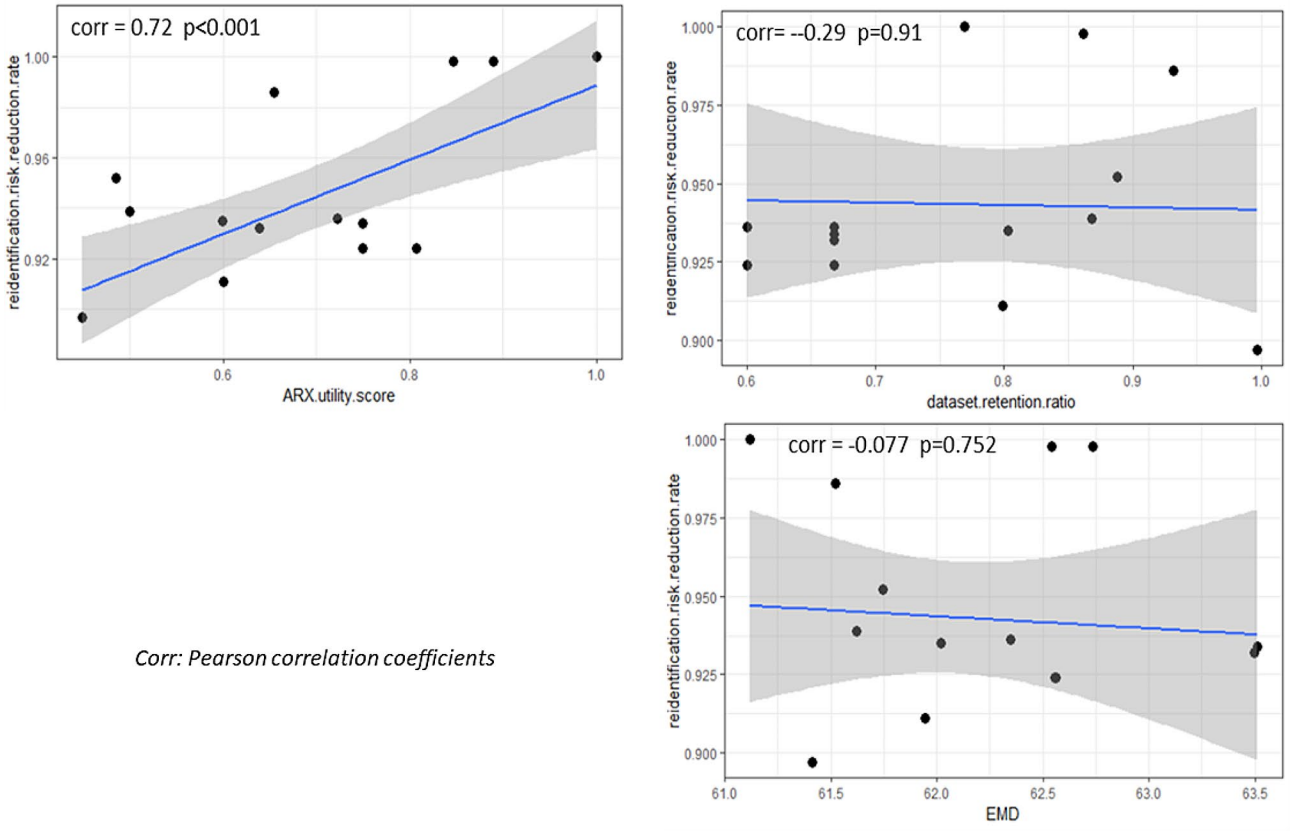


Fig. 4 The correlations between re-identification risk reduction and features of the de-identified datasets

Discussion

This study tested various de-identification strategies on a clinical dataset, adjusting the number and types of quasi-identifiers and sensitive information, and configuring K-anonymity, L-diversity, and T-closeness in diverse ways. It aimed to address gaps left by earlier studies that utilized simplistic data use cases and de-identification configurations [28, 34, 35].

The results indicated that de-identification led to the suppression of records and variables, precluding the replication of analyses performed on the original dataset. Consequently, logistic regression models for predicting

ED LOS yielded differing conclusions based on the de-identification approach, as illustrated in Fig. 3. This highlights the need for the evolution of privacy technologies that maintain data integrity. Additionally, it cautions data users about potential biases introduced when working with de-identified datasets.

The study found optimal data utility when only *sex* and *age* were classified as quasi-identifiers, maintaining all variables and losing only six records. This configuration also significantly reduced the baseline re-identification risk, albeit *sex* and *age* by themselves did not strongly individualize records. However, this configuration did not account for the additional re-identification risk posed by the *sending hospital* and *primary diagnosis*, both of which were considered the most identifying variables in the dataset (Table 2). To eliminate any alterations to *sex* and *age*—key variables for clinical research—we examined the impact of designating only the *sending hospital* and *primary diagnosis* as quasi-identifiers (dataset 19). This strategy greatly reduced the chance of re-identification but at a considerable cost to data utility, resulting in the loss of over half the dataset and two predictor variables: the *sending hospital type* and *high priority disease*.

Seeking a compromise, datasets 5–12 incorporated *sex*, *age*, and either *sending hospital* or *primary diagnosis* as quasi-identifiers. In this series, datasets 7 and 8 achieved zero re-identification risk post-de-identification but sacrificed nearly half of the predictor variables. Datasets 11 and 12, while managing to retain all records, were considered less favorable due to the loss of four predictor variables. Datasets 5 and 6 struck a more acceptable balance, offering substantial re-identification risk reduction, retaining over 78% of records, and sacrificing only one predictor variable. Although dataset 5 had marginally better scores for risk reduction and data utility, dataset 6 was preferred because it retained information on *high priority disease*, a key predictor of ED LOS.

In this study, three different data utility metrics were examined, but only the ARX utility score exhibited a statistically significant correlation with the re-identification risk reduction rate. The EMD and dataset retention ratio both showed minor negative correlations with re-identification risk reduction; however, these were not statistically significant. This could suggest that the structural aspects of a dataset may not alone be adequate for assessing its utility, although further studies with a broader array of datasets would be required to substantiate this preliminary indication.

The scope of this research was limited to a single use case, analyzing data obtained from one hospital. Moreover, the range of de-identification scenarios tested did not encompass the full spectrum of complex configurations that could be employed. Despite these constraints, the research offers valuable insights into the nuanced interplay between data de-identification processes and data utility. It contributes to the ongoing conversation about how to approach data privacy in a way that still enables effective data usage.

Conclusion

As health data analysis grows more critical, so does the imperative to devise effective methods for ensuring data privacy. While established guidelines [47] offer a foundation for the de-identification of datasets, crafting a dataset that maintains a high level of privacy without unduly compromising its utility remains a nuanced challenge. It demands a thorough grasp of the data's intended application. Incorporating input from data users during the de-identification process and considering the variety of potential data use cases could prove beneficial in finding a workable tradeoff between data privacy and utility.

Abbreviations

AUC	Area Under the receiver operating characteristic (ROC) Curve
DP	Differential Privacy
ED	Emergency Department
EMD	Earth Mover's Distance
HE	Homomorphic Encryption
ICD	International Classification of Disease
LOS	Length Of Stay
ROC	Receiver Operating Characteristic

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02545-9.

Supplementary Material 1

Acknowledgements

EI received a scholarship from the BK21 education program (Center for World-leading Human-care Nurse Leaders for the Future).

Author contributions

EI conducted data de-identification and data analysis. HK conceived the initial project idea and interpreted the results. EI and HK designed the study and wrote the manuscript. HL prepared the clinical data and analyzed the utility of the de-identified dataset. XJ and JK interpreted the analysis results and provided critical insights into the data de-identification approaches. All authors read and approved the final manuscript.

Funding

This study was supported in part by a research grant from the Korean Healthcare Bigdata showcase Project by the Korea Disease Control and Prevention Agency in the Republic of Korea (no.4800-4848-501). The funding body played no role in the design of the study and collection, analysis, interpretation of data, and writing the manuscript.

Data availability

The clinical dataset used in this study is not made available due to the sensitive nature of clinical data. However, de-identified analytic datasets are available upon reasonable request from the corresponding author and with permission of Seoul National University Hospital.

Declarations

Ethics approval and consent to participate

This study utilized retrospective EHR data and was approved by the Institutional Review Board of the Seoul National University Hospital Biomedical Research Institute (IRB approval No: H-2009-156-1159). In accordance with Article 16 of the Korean Bioethics Law, informed consent was waived by the IRB. All experiments were performed in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 1 June 2023 / Accepted: 21 May 2024
Published online: 30 May 2024

References

1. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25(1):37–43.
2. Gostin LO, Halabi SF, Wilson K. Health data and privacy in the digital era. JAMA. 2018;320(3):233–4.
3. Data Protection and Privacy Legislation Worldwide | UNCTAD. https://unctad.org/page/data-protection-and-privacy-legislation-worldwide. Accessed 6 Oct 2022.
4. 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. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#coveredentities (2022). Accessed 28 Mar 2024.
5. General Data Protection Regulation (GDPR). Article 32 GDPR(https://gdprhub.eu/index.php?title=Article_32_GDPR (2023). Accessed 4 Apr 2024.
6. Personal Information Protection Commission. Pseudonymization Guidelines. Korea;2024.
7. Thapa C, Camtepe S. Precision health data: requirements, challenges and existing techniques for data security and privacy. Comput Biol Med. 2021;129:104130.
8. Cho H, Simmons S, Kim R, Berger B. Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs. Cell Syst. 2020;10(5):408–16.e9.
9. Deldar F, Abadi M. Differentially private count queries over personalized-location trajectory databases. Data Brief. 2018;20:1510–4.
10. Venkatesaramani R, Wan Z, Malin BA, Vorobeychik Y. Enabling tradeoffs in privacy and utility in genomic data beacons and summary statistics. Genome Res. 2023;33(7):1113–23.
11. Xiong L, Post A, Jiang X, Ohno-Mochado L. New Methods to Protect Privacy When Using Patient Health Data to Compare Treatments. 2021.
12. Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, Ienca M, Fellay J, Vayena E, et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Res. 2021;23(2):e25120.
13. Bataa M, Song S, Park K, Kim M, Cheon JH, Kim S. Finding highly similar regions of genomic sequences through homomorphic encryption. J Comput Biol. 2024;31(3):197–212.
14. Kim D, Son Y, Kim D, Kim A, Hong S, Cheon JH. Privacy-preserving approximate GWAS computation based on homomorphic encryption. BMC Med Genom. 2020;13:1–12.
15. Rovida L, Leporati A. Encrypted image classification with low memory footprint using fully homomorphic encryption. Cryptology ePrint Archive; 2024.
16. Acar A, Aksu H, Uluagac AS, Conti M. A survey on homomorphic encryption schemes: theory and implementation. ACM Comput Surv (Csur). 2018;51(4):1–35.
17. Kuo T-T, Kim H-E, Ohno-Machado L. Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc. 2017;24(6):1211–20.
18. Zhang F, Zhang Y, Ji S, Han Z. Secure and decentralized Federated Learning Framework with Non-IID Data based on Blockchain. Heliyon. 2024.
19. Wu C, Tang YM, Kuo WT, Yip HT, Chau KY. Healthcare 5.0: a secure and distributed network for system informatics in medical surgery. Int J Med Informatics. 2024:105415.
20. Ali A, Al-Rimy BAS, Tin TT, Altamimi SN, Qasem SN, Saeed F. Empowering Precision Medicine: Unlocking Revolutionary insights through Blockchain-enabled Federated Learning and Electronic Medical Records. Sensors. 2023;23(17):7476.
21. Chukwu E, Garg L. A systematic review of blockchain in healthcare: frameworks, prototypes, and implementations. Ieee Access. 2020;8:21196–214.
22. Fan C, Ghaemi S, Khazaei H, Musilek P. Performance evaluation of blockchain systems: a systematic survey. IEEE Access. 2020;8:126927–50.
23. Thantilage RD, Le-Khac N-A, Kechadi M-T. Healthcare data security and privacy in Data Warehouse architectures. Inf Med Unlocked. 2023:101270.
24. Tandon A, Dhir A, Islam AN, Mäntymäki M. Blockchain in healthcare: a systematic literature review, synthesizing framework and future research agenda. Comput Ind. 2020;122:103290.
25. Ahmed T, Aziz MMA, Mohammed N. De-identification of electronic health record using neural network. Sci Rep. 2020;10(1):18600.
26. Ahmed T, Aziz MMA, Mohammed N, Jiang X, editors. Privacy preserving neural networks for electronic health records de-identification. Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics; 2021.
27. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzziness Knowledge-Based Syst. 2002;10(05):571–88.
28. Jeon S, Seo J, Kim S, Lee J, Kim J-H, Sohn JW, et al. Proposal and assessment of a de-identification strategy to enhance anonymity of the observational medical outcomes partnership common data model (OMOP-CDM) in a public cloud-computing environment: anonymization of medical data using privacy models. J Med Internet Res. 2020;22(11):e19597.
29. Personal Information Protection Commission. uidelines for Personal Information De-identification Measures. 2016.
30. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. Acm Trans Knowl Discovery data (tkdd). 2007;1(1):3–es.
31. Li N, Li T, Venkatasubramanian S, editors. t-closeness: Privacy beyond k-anonymity and l-diversity. 2007 IEEE 23rd international conference on data engineering; 2006: IEEE.
32. Tomashchuk O, Van Landuyt D, Pletea D, Wuyts K, Joosen W, editors. A data utility-driven benchmark for de-identification methods. Trust, Privacy and Security in Digital Business: 16th International Conference, TrustBus 2019, Linz, Austria, August 26–29, 2019, Proceedings 16; 2019: Springer.
33. Brickell J, Shmatikov V, editors. The cost of privacy: destruction of data-mining utility in anonymized data publishing. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining; 2008.
34. Wu L, He H, Zaïane OR, editors. Utility of privacy preservation for health data publishing. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; 2013: IEEE.
35. Li T, Li N, editors. On the tradeoff between privacy and utility in data publishing. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining; 2009.
36. Karagiannis S, Ntantogian C, Magkos E, Tsohou A, Ribeiro LL. Mastering data privacy: leveraging K-anonymity for robust health data sharing. Int J Inf Secur. 2024:1–13.
37. Zamani A, Oechtering TJ, Skoglund M. On the privacy-utility trade-off with and without direct access to the private data. IEEE Trans Inf Theory. 2023.
38. Baek S-M, Seo D-W, Kim Y-J, Jeong J, Kang H, Han KS, et al. Analysis of emergency department length of stay in patient with severe illness code. J Korean Soc Emerg Med. 2020;31(5):518–25.
39. Laam LA, Wary AA, Strony RS, Fitzpatrick MH, Kraus CK. Quantifying the impact of patient boarding on emergency department length of stay: all admitted patients are negatively affected by boarding. J Am Coll Emerg Physicians Open. 2021;2(2):e12401.
40. Otto R, Blaschke S, Schirrmeister W, Drynda S, Walcher F, Greiner F. Length of stay as quality indicator in emergency departments: analysis of determinants in the German Emergency Department Data Registry (AKTIN registry). Intern Emerg Med. 2022;17(4):1199–209.
41. National Emergency Medical Center: Statistical yearbook of National Emergency Department Information System. https://www.e-gen.or.kr/nemc/statistics_annual_report.do?%20brdclscd=02 (2022). Accessed 7 Oct 2022.

42. Chang Y-H, Shih H-M, Chen C-Y, Chen W-K, Huang F-W, Muo C-H. Association of sudden in-hospital cardiac arrest with emergency department crowding. Resuscitation. 2019;138:106–9.
43. Kim J-s, Bae H-J, Sohn CH, Cho S-E, Hwang J, Kim WY, et al. Maximum emergency department overcrowding is correlated with occurrence of unexpected cardiac arrest. Crit Care. 2020;24:1–8.
44. Lee H, Lee S, Kim H. Factors affecting the length of stay in the emergency department for critically ill patients transferred to regional emergency medical center. Nurs Open. 2023;10(5):3220–31.
45. World Health Organization(WHO). International Statistical Classification of Diseases and Related Health Problems(ICD). https://www.who.int/standards/classifications/classification-of-diseases/1 (2019). Accessed 11 Oct, 2022.
46. Eicher J, Kuhn KA, Prasser F. An experimental comparison of quality models for health data de-identification. MEDINFO 2017: Precision Healthcare through Informatics: IOS; 2017. p. 704–8.
47. Jakob CE, Kohlmayer F, Meurers T, Vehreschild JJ, Prasser F. Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Sci data. 2020;7(1):435.
48. Meurers T, Bild R, Do K-M, Prasser F. A scalable software solution for anonymizing high-dimensional biomedical data. GigaScience. 2021;10(10):giab068.
49. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA, editors. Arx-a comprehensive tool for anonymizing biomedical data. AMIA Annual Symposium Proceedings; 2014: American Medical Informatics Association.
50. ARX Configuration. n.d. https://arx.deidentifier.org/anonymization-tool/configuration/. Accessed 4 Apr 2024.
51. Pele O, Werman M, editors. Fast and robust earth mover's distances. 2009 IEEE 12th international conference on computer vision; 2009: IEEE.
52. Gart JJ. The comparison of proportions: a review of significance tests, confidence intervals and adjustments for stratification. Revue de l'Institut International de Statistique; 1971. pp. 148–69.
53. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988:837–45.
54. R Core Team. R: a language and environment for statistical. Version 4.0.4. Vienna. Austria: R Foundation for Statistical Computing; 2021.

Publisher's Note