İSTANBUL UNIVERSITY CERRAHPAŞA

ORIGINAL ARTICLE
Electrica 2023; 23(3): 658-665

Collection of an e-Health Dataset and Anonymization with Privacy-Preserving Data Publishing Algorithms

Burak Cem Kara[1,2], Can Eyüpoğlu[1], Serkan Uysal[2], Selim Bayraklı[1]

[1]Department of Computer Engineering, National Defence University, Turkish Air Force Academy, İstanbul, Turkey
[2]Department of Computer Engineering, National Defence University, Atatürk Strategic Studies and Graduate Institute, İstanbul, Turkey

Cite this article as: B.C. Kara, C. Eyüpoğlu, S. Uysal and S. Bayraklı, "Collection of an e-health dataset and anonymization with privacy-preserving data publishing algorithms," *Electrica*, 23(3), 658-665, 2023.

ABSTRACT

The healthcare field remains one of the most important social and economic challenges worldwide, demanding new and more advanced solutions from science and technology. The developments in Industry 4.0 have brought tremendous improvements in the healthcare industry such as better quality of treatment, improved communication, remote monitoring, and lower cost. Therefore, sharing health data with healthcare providers is important for the continuation of such improvements. In general, health data contain sensitive information about individuals. Therefore, in accordance with privacy regulations and ethical requirements, it is essential to protect patients' privacy before sharing data for medical research. In this study, a new dataset is created with the data collected from e-Health platforms where patients and doctors meet. Some *k*-anonymization-based algorithms, which are frequently used in the literature, have been applied to this created health dataset. In order to reveal the success of the algorithms applied to the health dataset, various information metrics have been used and the results obtained from the experiments are presented in detail.

***Index Terms*—**Data anonymization, e-health, healthcare, k-anonymity, privacy-preserving data publishing.

I. INTRODUCTION

Today, people who actively use technology products and are referred to as users of information services publish many sensitive personal data in their online transactions, including identity information such as name, surname, gender, date of birth, marital status, as well as disease information, daily activities, private interests, etc. In this way, the presence, collection, analysis, and use of personal data in the open network create potential opportunities for both positive and negative purposes for many individuals or institutions [1]. One of the places where this situation is commonly seen is e-Health platforms where patients and doctors meet. When these types of platforms are examined, basic methods such as generalization and masking are used to ensure the confidentiality of data that would reveal individuals' identity information. Although the information that would fully identify a person is not shared or attempted to be blocked on these sites where sensitive data are shared, it is not difficult for attackers to take advantage of this situation and reach the person. In order to alleviate individuals' security and privacy concerns related to their personal information, different protection methods have been developed in the literature [2].

In this study, the health dataset created with the data collected on such e-Health platforms, which people think that they provide complete data privacy on personal privacy by preventing the information that fully identifies them with some simple anonymization methods, is presented. Besides, some *k*-anonymization-based algorithms, which are frequently used in the literature, have been applied to all kinds of full identifiers, quasi-identifiers, and sensitive and non-sensitive information that patients shared while seeking a cure for their disease through such platforms. It should be noted that using this type of health data where patient stories are shared, doctors make inferences from the shared data and diagnose. Therefore, the success of the algorithms has been tested, taking into account that the useful information shared on the platform is not subject to over-anonymization.

Corresponding author:
Burak Cem Kara

E-mail:
burakcemkara@gmail.com.

Received: March 20, 2023

Accepted: June 12, 2023

Publication Date: August 1, 2023

DOI: 10.5152/electrica.2023.23042

The contributions of this study are listed as follows:

- A new e-Health dataset obtained from all kinds of full identifiers, quasi-identifiers, and sensitive and non-sensitive information shared while seeking cures for diseases on e-Health platforms is created and presented in the literature.
- The created health dataset has been used as an experimental dataset for the first time in this study, and the success of the algorithms has been demonstrated in terms of various information metrics by applying *k*-anonymization-based algorithms to the dataset.
- Three basic anonymization algorithms which are Mondrian, Datafly, and Top-Down have been tested on this real-world dataset.

The organization of the rest of the paper is as follows: In Section II, background and related work about study topic are given. In Section III, the dataset and algorithms used in the study are introduced. In Section IV, the information metrics utilized are mentioned and the results of the experiments conducted are demonstrated. Finally, the study is concluded in Section V.

II. BACKGROUND AND RELATED WORK

k-Anonymity algorithm provides privacy preserving by using the anonymization methods which are available in the literature such as generalization, masking, slice, permutation, and perturbation. Among these anonymization methods, generalization is still a popular methodology. The main idea of the generalization method shown in Fig. 1 is to ensure that the adversary cannot disclose the sensitive information of individuals using quasi-identifiers in datasets. However, generalizations lead to the loss of information. Therefore, overgeneralization should be avoided as long as privacy requirements are met. The existing generalization algorithms have adopted two different approaches. These are called local and global generalizations, also referred to as local and global recoding [3]. Local generalization-based algorithms offer a superior data utility compared to global generalization-based algorithms. The purpose of data publication is to use published data, so many researchers are concerned with the quality of generalized data. Datafly [4], Incognito [5], Mondrian [6], Basic Mondrian [6], Top-Down greedy data anonymization [7], and clustering-based *k*-anonymization algorithms [8, 9] are the most widely used methods.

The Incognito generates all possible *k*-anonymous full-domain generalizations of a given dataset *T* while allowing for optional tuple suppression. The algorithm starts by checking single-attribute subsets of the quasi-identifier and iterates by checking *k*-anonymity with respect to increasingly large subsets [5]. Clustering-based *k*-anonymization is an increasingly popular algorithm in the literature. Byun et al. [8] proposed the greedy *k*-member clustering algorithm for *k*-anonymization. In another study [9], a different clustering-based method is presented. This method, unlike the study [8], tries to construct all clusters at once and is also more robust to outliers. The Basic Mondrian is a multidimensional *k*-anonymity algorithm and an extended version of Mondrian algorithm [6]. It provides a strong theoretical basis for constructing a multidimensional model of *k*-anonymity. LeFevre et al. [6] proposed a different version of Mondrian called Basic Mondrian that supports both categorical and numeric attributes to solve this problem. The Basic Mondrian algorithm follows a hierarchy tree to generalize the categorical attributes. Overall, it has a good performance in generalizing categorical attributes [10].

In recent years, many algorithms based on *k*-anonymity have been developed [11, 12, 13]. Among these works, Kacha et al. [11] proposed a new *k*-anonymity approach based on the black hole algorithm. This approach, called *k*-anonymity based on black hole algorithm (KAB), relies on the use of the black hole optimization (BHO) algorithm during the *k*-anonymization process. KAB aims to provide a higher level of protection compared to traditional *k*-anonymization methods. To achieve this goal, the KAB approach uses not only traditional approaches but also the accuracy and scalability of the BHO algorithm to ensure privacy. The study demonstrates the success of the KAB method by testing it on various datasets. Kiran and Shirisha [12] propose a new model based on *k*-anonymization that uses a combination of perturbation techniques, including adding noise and shuffling data. This model is tested on several datasets and compared to other *k*-anonymization methods, demonstrating its effectiveness in preserving privacy while maintaining data utility. In the work of Mahanan et al. [13], a privacy-preserving algorithm based on *k*-anonymity is presented. The algorithm utilizes a heuristic method to select the best quasi-identifier attributes to ensure that the *k*-anonymity requirement is satisfied. Then, the algorithm applies a data perturbation technique to the selected quasi-identifier attributes to protect sensitive information. The authors evaluate the proposed algorithm on several real-world datasets and show that it achieves a good balance between data privacy and data utility, outperforming existing *k*-anonymity-based algorithms.

III. MATERIALS AND METHODS

A. Dataset

Combined with the convenience and comfort provided by e-Health platforms, it is not difficult to predict that patients will prefer e-Health platforms over visiting healthcare centers. In addition to the benefits that this new situation brings to individuals, it also creates many problems regarding data privacy and confidentiality. In this study, some of the leading algorithms in the literature have been implemented on the health dataset created within the scope of the study. The health dataset has been collected from e-Health platforms where patients and doctors meet.

Attributes in a dataset to be anonymized are examined in four different classes: identifiers, quasi-identifiers, sensitive attributes, and non-sensitive attributes [14]. Identifier is information such as name-surname, passport number, and phone number that can be used to

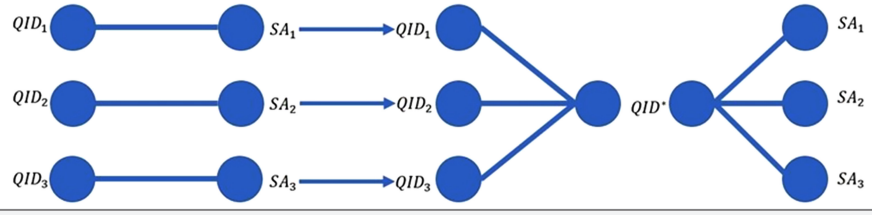

Fig. 1. Generalization methodology.

uniquely identify a person. If one of these information is available in a dataset, it is quite possible for a person to be exposed. Quasi-identifier is a set of attributes that alone is not sufficient to identify an individual but will enable the individual to be disclosed if linked to some external dataset or matched to more than one record. Age, gender, ethnicity, and date of birth can be given as examples for quasi-identifiers. Sensitive attributes are attributes that are not wanted to be known by third parties. An individual's salary information and illness status are examples of sensitive attributes. Non-sensitive attributes are the attributes that will not violate the user's privacy in case of disclosure and fall outside of the above classifications [1]. The created e-Health dataset contains a total number of 1086 records. There is a total of eight quasi-identifier attributes, one of which is numeric and seven of which is categorical. The diagnosis attribute is used as sensitive attributes. A detailed representation of the e-Health dataset is available in Table I.

B. Mondrian

Mondrian is an anonymization model that uses the *k*-dimensional tree (KD-Tree) to divide the data space multidimensionally and ensures confidentiality by generalizing on the resulting equivalence classes. The KD-tree method is an area partitioning tree that divides data points into some lower dimensional areas according to their projections [15]. In Mondrian, the purpose of using KD-Tree is to recursively divide data fields into binary subgroups [16]. In the KD-tree phase, a dimension is first selected, and the frequency of the data in the selected dimension is calculated. Then, the median of these frequencies is found and accepted as the division value. The data are partitioned according to the split value, and then two subsets are obtained. These operations are repeated until there are no data left for splitting. In the second phase, each of these small subsets is generalized. Thus, an anonymized dataset is obtained. The pseudo-code of the Mondrian algorithm is given in Algorithm 1.

Algorithm 1: Mondrian	
1:	**Function** *Mondrian* (*data, k*)
2:	If $k \leq \|data\| < l$ *then*
3:	*return generalize* (*data*)
4:	**else**
5:	*dim = select_dimension* (*data*)
6:	*fs = frequency_set* (*data,dim*)
7:	*splitVal = find_median* (*fs*)
8:	*lhs = partition* $\in$ *data : partition.dim* $\leq$ *splitVal*
9:	*rhs = partition* $\in$ *data : partition.dim > splitVal*
10:	*return Mondrian* (*lhs, k*) *U Mondrian* (*rhs, k*)
11:	**endif**
12:	**endfunction**

C. Datafly

Datafly is a local optimum algorithm developed to ensure *k*-anonymity, where the search space is the whole lattice. The Datafly algorithm works on the assumption that the best solutions are those obtained after generalizing the variables to the most specific values (unique items). Datafly utilizes a greedy algorithm to explore the domain generalization hierarchy [17]. In this algorithm, generalization and masking are used as anonymization methods. Datafly is a heuristic algorithm that performs one-dimensional full-domain generalization. The pseudo-code of the Datafly algorithm is given in Algorithm 2.

TABLE I. SUMMARY OF THE E-HEALTH DATASET

Attribute	Attribute Type	Distinct Values	Domain
Age	Continuous	73	[17–90]
Gender	Nominal	2	Male, female.
Education	Nominal	7	Ph.D., post-graduate, undergraduate, high school, middle school, elementary school, out-of-school.
Marital status	Nominal	3	Never married, married, divorced.
City	Nominal	8	Balikesir, Istanbul, Ankara, Bursa, Izmir, Konya, Kars, Diyarbakir
Symptom	Nominal	52	Prolactin-high, vaginal-bleeding, urination-burning, stomach-ache-constipation-fever, high-pulse, blood-pressure, panic-attack, heart-burn, groin-pain, bladder-pain, hip-and-leg-pain, bone-swelling, nausea-vomiting, dizziness, armpit-swelling, shortness-of-breath, chest-pain, lip-swelling, hypoglycemia, eyelid-whitening, constipation, prolactin-high, testicular-instability, eye- tremor, heart-palpitations.
Disease-category	Nominal	14	Gynecology, internal-medicine, physical-medicine-and-rehabilitation, orthopedics-and-traumatology, urology, plastic-reconstructive-and-aesthetic-surgery, brain-and-nerve-surgery, neurology, psychiatry, dermatology, cardiology, general-surgery, ear-nose-throat-diseases, eye-diseases.
Diagnosis	Nominal	55	Subclinical-hypothyroidism, tamoxifen, infection, cyst, hypotension-secondary, anxiety, neural-therapy, helicobacter-gastritis, waist-area-tumor, prostate, piriformis-syndrome, bone-lesion, adipose-tissue, vertigo, fibromyalgia, anxiety- disorder, hair-roll, aortic-coarctation, asthma, menopause, hormone-imbalance, juvenile-gynecomastia, meniscus-tear, insulin-resistance, reflux, varicocele, vitiligo.

Algorithm 2: Datafly	
1:	**function** Datafly algorithm(*data, k*):
2:	**for each** column *i* in *data*:
3:	Compute the mean *M_i* of the values in column *i*
4:	Compute the standard deviation *S_i* of the values in column *i*
5:	**for each** row *j* in *data*:
6:	Compute the Datafly value *D_i, j* of the value in column *i*, row *j* as follows:
7:	*D_i, j* = (*data_i, j* - *M_i*) / *k* * *S_i*
8:	**return** Datafly *data*
9:	**end function**

D. Top-Down

The Top-Down greedy data anonymization algorithm is an iterative method that changes from the topmost domain values through the taxonomy trees of the attributes [18]. The table is partitioned into sections iteratively. If every subset is more local, an equivalence class set is partitioned into subsets. Moreover, they are likely to be partitioned into smaller groups, which reduces the weighted precision penalty. After partitioning, groups smaller than *k* are united to provide the *k*-anonymity requirement [7]. The basic Top-Down algorithm is given in Algorithm 3.

Algorithm 3: Top-Down	
1:	**function** Top_Down algorithm(*problem*):
2:	**if** *problem* is small enough to solve directly:
3:	**return** solution to *problem*
4:	**else:**
5:	Divide *problem* into smaller *subproblems*
6:	*Solution_1* = Top_Down algorithm(*Subproblem_1*)
7:	Solution_2 = Top_Down algorithm(*Subproblem_2*)
8:	...
9:	*Solution_n* = Top_Down algorithm(*Subproblem_n*)
10:	**return** Combine(*Solution_1*, *Solution_2*, ..., *Solution_n*)
11:	**end if**
12:	**end function**

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the experimental evaluation of the algorithms applied to the e-Health dataset is presented. A notebook with 16 GB RAM and ninth-generation i7 processor was used in the analysis studies. PyCharm was used as the integrated development environment for the experiments. To measure the success of the algorithms, discernibility metric (DM), normalized average equivalence class size (Cavg), normalized certainty penalty (NCP), and run time metrics were used. The e-Health dataset was used for the experiments.

A. Discernibility Metric

The DM is used to quantify the level of indistinguishability of a record from others, with each record within the equivalence class being assigned equal penalty scoring [19, 20]. The penalty value for a record belonging to the corresponding equivalence class is determined by the size of the equivalence class *T* in the *Z* table, denoted as $|T|$. The DM can be expressed as:

$$DM(Z) = \sum_{\text{EquivClassesT}} |T|^2 \quad (1)$$

Fig. 2 presents a comparison of the DM results of the anonymization algorithms used. The algorithm with a lower DM score is more successful. As the *k* value increases, it is seen that the DM values of all algorithms increase. Among the algorithms applied to the e-Health dataset, it is observed that the Top-Down algorithm achieves the most successful results at low *k* values, while the Mondrian algorithm is more successful as the *k* values increase. The Datafly algorithm obtained the most unsuccessful results.

B. The Normalized Average Equivalence Class Size

The Cavg metric was originally introduced in [16] and later utilized as an information metric in [7]. It measures the degree to which the *k*-anonymization model applied to the dataset results in data loss by computing the average size of the generated equivalence classes [20]. The Cavg score of an anonymized table is determined using the following formula [11]:

$$Cavg = \frac{total\_records}{total\_equiv\_classes} / (k) \quad (2)$$

In this equation, *total_records* represent the number of records in the dataset, *total_equiv_classes* denote the number of created equivalence classes, and *k* denotes the privacy level. The optimal score for the Cavg metric is 1 [11].

In Fig. 3, the performances of the algorithms against the Cavg metric are presented. Cavg measures data utility based on the dimensions of the equivalence classes. In terms of the Cavg metric, 1 value is a baseline, and algorithms close to this value performed better. When the graph is examined, it is observed that the Top-Down algorithm is more successful than other algorithms at low *k* values. At high *k* values, it is seen that the Mondrian and Top-Down algorithms obtain close results. The Datafly algorithm gives the worst result for all *k* values.

C. Normalized Certainty Penalty

The NCP metric is utilized to assess the accuracy loss in defining equivalence classes [21]. NCP works by measuring the information loss of a complete partition by summing up all the equivalence classes in each group. NCP score ranges from 1 to 0, where a score closer to 0 indicates little or no information loss, while a value near 1 suggests a significant loss of information. If Q_i represents a numeric or categorical attribute, *d* represents the number of attributes in a dataset, and w_i represents weights. The NCP score of the *T* is calculated as follows:

$$NCP(T) = \sum_{i=1}^{d} w_i \cdot NCP_{Q_i}(T) \quad (3)$$

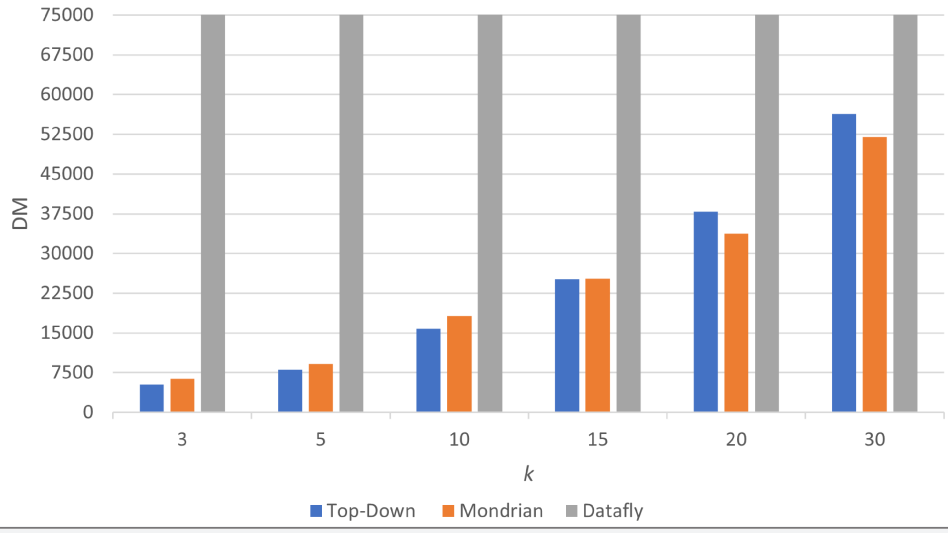

Fig. 2. Comparison of discernibility metric results.

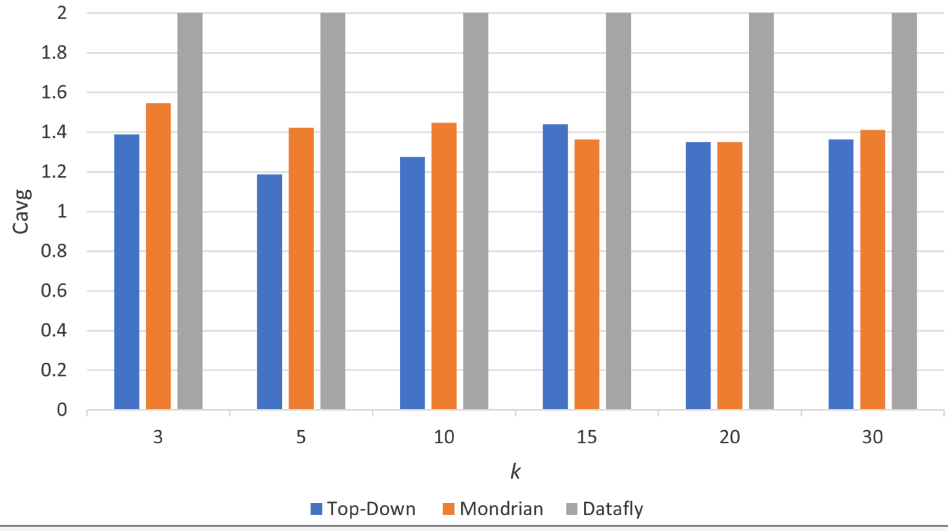

Fig. 3. Comparison of normalized average equivalence class size results.

In Fig. 4, the results of the NCP metric used to measure the loss of information in the anonymized dataset are presented. The algorithms with lower NCP scores in the graph performed better. According to these results, as the *k* value increases, the NCP score of all algorithms gradually increases. Because when the value of *k* increases, all algorithms distort more information to achieve *k*-anonymity. In this case, it negatively affects and increases the NCP value. According to these results, it is observed that the Mondrian algorithm obtained the best results for all *k* values.

D. Run Time

Run time refers to the time elapsed between the execution of a program and its termination. In this study, the run times of the algorithms utilized are compared, and the results are shown in Fig. 5.

As seen in Fig. 5, the Top-Down algorithm has a worse running time than the other two algorithms. Among these three algorithms, the best running time belongs to the Mondrian algorithm. Datafly is more successful than Top-Down in terms of run time.

E. Discussion

In this study, the performances of three anonymization algorithms, Mondrian, Top-Down, and Datafly, are evaluated using four different metrics. According to the results of the Cavg metric, the Top-Down algorithm performs better than the other algorithms at low *k* values, while the Mondrian and Top-Down algorithms have similar results at high *k* values. On the other hand, the Datafly algorithm gives the worst results for all *k* values. When the results of the NCP metric are evaluated, it is seen that the Mondrian algorithm obtains the best results for all *k* values. In the DM, the Top-Down algorithm achieves the most successful results at low *k* values, while the Mondrian algorithm is more successful as *k* values increase. The Datafly algorithm attains the most unsuccessful results. Finally, run time is evaluated as a metric, and it is seen that the Mondrian algorithm has the best

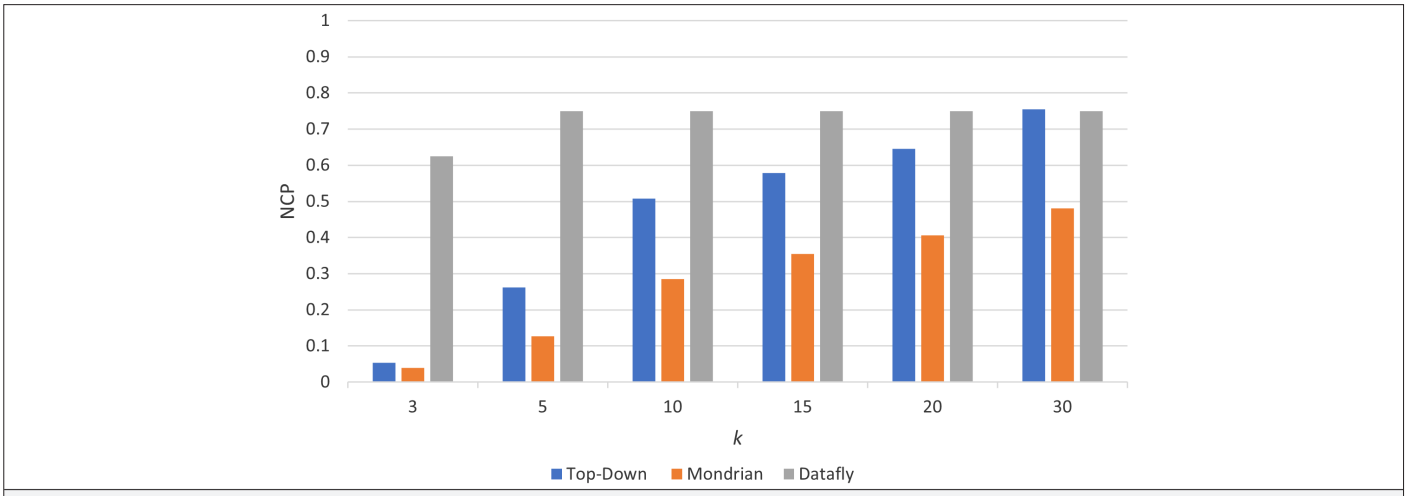

Fig. 4. Comparison of normalized certainty penalty results.

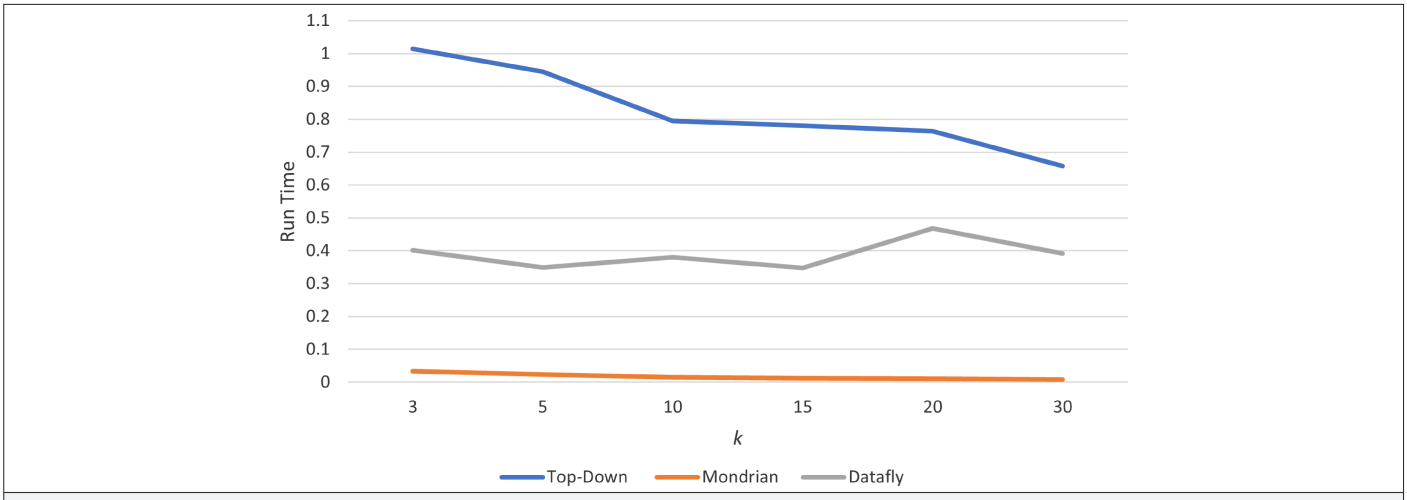

Fig. 5. Comparison of run time results.

running time among the three algorithms. The Top-Down algorithm has a worse running time than the other two algorithms. In conclusion, the experimental results show that the Mondrian algorithm outperforms the other two algorithms in terms of NCP and run time metrics. The Top-Down algorithm is more successful at low *k* values according to the Cavg and DM. However, the Datafly algorithm is the least successful algorithm among the evaluated algorithms.

V. CONCLUSION AND FUTURE WORK

The issue of protecting the personal information of patients on e-Health platforms needs a more secure structure against the ever-evolving and changing attack models. With the coronavirus disease 2019 pandemic that has emerged in recent years, there is a serious digitalization transformation in the health sector, as in many sectors. Due to the pandemic, patients have had some concerns about going to hospitals or health centers to be examined. Therefore, it has become a common situation for them to prefer digital environments where they can communicate with doctors safely and easily without endangering their health. In this study, first, a unique dataset has been created with the data collected from e-Health platforms where patients and doctors meet within all these predictions. Mondrian, Datafly, and Top-Down algorithms have been applied to this created dataset to ensure data privacy by keeping data availability at the highest level. According to the experimental results, the Top-Down has obtained more successful results regarding metrics that measure the quality of equivalence classes such as DM and Cavg. Mondrian showed a clear superiority in terms of the NCP metric that measures the data utility. In regard to running time, the algorithms are ranked as Mondrian, Datafly, and Top-Down from the best to the worst.

For future studies, new *k*-anonymization-based models can be developed and applied to the dataset proposed in this study. Additionally, the effectiveness of the anonymization methods applied to the dataset can be tested not only with *k*-anonymization algorithms but also with models using different privacy techniques, such as differential privacy.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – C.E., B.C.K.; Design – C.E., B.C.K, S.B.; Supervision – C.E., S.B.; Materials – B.C.K., S.U.; Data Collection and/or Processing – S.U., B.C.K.; Analysis and/or Interpretation – B.C.K., S.U., C.E., S.B.; Literature Review – B.C.K., S.U.; Writing – B.C.K.; Critical Review – C.E., B.C.K.

Declaration of Interests: The authors declare that they have no competing interest.

Funding: This study received no funding.

REFERENCES

1. B. C. Kara and C. Eyupoglu, "Privacy and security problems in healthcare 4.0," 4th International Symposium on Multidisciplinary Studies and Innovative Technologies. 2020, pp. 1–12, Online, IEEE.
2. B. C. Kara and C. Eyupoglu, "Anonymization methods for privacy-preserving data publishing," *Smart Appl. with Adv. Mach. Learn. Human-Centred Probl. Des. Eng. Cyber-Physical Syst. Crit. Infrastructures*, Vol. 1, 2023, pp. 145–159. [CrossRef]
3. K. Arava and S. Lingamgunta, "Adaptive k-anonymity approach for privacy preserving in cloud," *Arab. J. Sci. Eng.*, vol. 45, no. 4, pp. 2425–2432, 2020. [CrossRef]
4. L. Sweeney, "Datafly: A system for providing anonymity in medical data," in *Database Security XI: Status and Prospects*, 1998, pp. 356–381, Springer. [CrossRef]
5. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," International Conference on Management of Data. 2005, pp. 49–60. [CrossRef]
6. K. LeFevre, D. J. Dewitt, and R. Ramakrishnan, "Workload-aware anonymization," International Conference on Knowledge Discovery and Data Mining. 2006, pp. 277–286. [CrossRef]
7. J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu, "Utility-based anonymization using local recoding," International Conference on Knowledge Discovery and Data Mining. Vol. 2006, 2006, pp. 785–790. [CrossRef]
8. J. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," 12th International Conference on Database Systems for Advanced Applications. 2007, pp. 188–200, Thailand.
9. J. L. Lin and M. C. Wei, "An efficient clustering method for k-anonymization," International Conference Proceedings Series, Vol. 331, 2008, pp. 46–50, France.
10. K. C. Liu, C. W. Kuo, W. C. Liao, and P. C. Wang, "Optimized data deidentification using multidimensional k-anonymity," International Conference on Trust, Security and Privacy in Computing and Communications. no. 3, 2018, pp. 1610–1614.
11. L. Kacha, A. Zitouni, and M. Djoudi, "KAB: A new k-anonymity approach based on black hole algorithm," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4075–4088, 2022. [CrossRef]
12. A. Kiran, and N. Shirisha, "K-anonymization approach for privacy preservation using data perturbation techniques in data mining," *Mater. Today Proc.*, vol. 64, pp. 578–584, 2022. [CrossRef]
13. W. Mahanan, W. A. Chaovalitwongse, and J. Natwichai, "Data privacy preservation algorithm with k-anonymity," *World Wide Web*, vol. 24, no. 5, pp. 1551–1561, 2021. [CrossRef]
14. C. Eyupoglu, M. A. Aydin, A. H. Zaim, and A. Sertbas, "An efficient big data anonymization algorithm based on chaos and perturbation techniques," *Entropy (Basel)*, vol. 20, no. 5, pp. 1–18, 2018. [CrossRef]
15. J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, 509–517, 1975. [CrossRef]
16. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional kanonymity," Int. Conf. Data Eng. Vol. 2006, 2006, p. 25.
17. A. K. Pal, *Achieving k-Anonymity Using Full Domain Generalization* (PhD Thesis). 2014, National Institute of Technology Rourkela, India.
18. S. Kavitha, S. Eswaran, and P. R. Vadhana, "A survey on k-anonymity generalization algorithms," *Int. J. Adv. Res Comput. Commun. Eng.*, pp. 8471–8474, 2014, Vol. 3, Issue 11.
19. R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," 21st International Conference On Data Engineering. 2005, pp. 217–228. [CrossRef]
20. J. Li, R. C.-W. Wong, A. W. C. Fu, and J. Pei, "Anonymization by local recoding in data with attribute hierarchical taxonomies," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1181–1194, 2008. [CrossRef]
21. M. Terrovitis, N. Mamoulis, and P. Kalnis, "Local and global recoding methods for anonymizing set-valued data," *VLDB J.*, vol. 20, no. 1, pp. 83–106, 2011. [CrossRef]

Burak Cem Kara is a Ph.D. student and research assistant in the Department of Computer Engineering, Air Force Academy, National Defence University, Istanbul, Turkey, where he has given lectures on an introduction to computer science and computer programming since 2019. He has published studies at both national and international conferences. Moreover, he has published some articles in national magazines. His research interests focus on big data, the Internet of Things, and personal and corporate information security and related fields. He currently not only gives lectures but also works as a manager in IT Department at the Air Force Academy.

Can Eyupoglu received the B.Sc. degree with high honor in Computer Engineering and Minor degree in Electronics Engineering from Istanbul Kültür University, Turkey in 2012, the M.Sc. and Ph.D. degrees with high honor in Computer Engineering from Istanbul University in 2014 and 2018, respectively. He is currently an Associate Professor and Head of Department in the Computer Engineering Department, Air Force Academy, National Defence University, Istanbul, Turkey. His current research interest includes data privacy, machine learning, data mining, bioinformatics, and image processing. He has published about 50 papers in various esteemed journals and conferences and has been serving as a member of the reviewer board in nearly 30 prestigious academic journals.

Serkan Uysal received the associate degree from Balıkesir University in 2015, B.Sc. degree in Computer Engineering from Kütahya Dumlupınar University in 2019, and the M.Sc degree in Cyber Security from National Defence University Atatürk Strategic Studies and Graduate Institute in 2023. He has been working at the Air Force Academy since 2020.

Selim Bayraklı received his B.Sc. degree with high honor, M.Sc. degree with high honor, and Ph.D. degree in Computer Engineering from Maltepe University, Turkey in 2006, 2008, and 2013, respectively. He is currently an Assistant Professor in Computer Engineering Department, Air Force Academy, National Defence University, Istanbul, Turkey. His current research interests include wireless sensor networks, Internet of Things, heuristic algorithms, data mining, and big data.