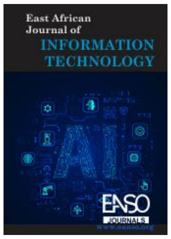

East African Journal of Information Technology
eajit.eanso.org
Volume 5, Issue 1, 2022
Print ISSN: 2707-5346 | Online ISSN: 2707-5354
Title DOI: https://doi.org/10.37284/2707-5354

EAST AFRICAN
NATURE &
SCIENCE
ORGANIZATION

Original Article

Assessing Identity Disclosure Risk in the Absence of Identified Datasets in the Public Domain

Peter N. Muturi[1*], *Dr. Andrew M. Kahonge, PhD*[1], *& Dr. Christopher Kipchumba Chepken, PhD*[1]

[1] University of Nairobi, P. O. Box 30197, GPO, Nairobi, Kenya.
* Correspondence ORCID ID: https://orcid.org/000-0002-8080-5110; email: pmuturi@mmu.ac.ke.

Article DOI: https://doi.org/10.37284/eajit.5.1.773

Date Published:

06 August 2022

Keywords:

Anonymisation,
De-Identification,
Re-Identification,
Privacy,
Data Release,
Data Analytics,
Analytical Utility

ABSTRACT

Data release is essential in supporting data analytics and secondary data analyses. However, data curators need to ensure the released datasets preserve data subjects' privacy and retain analytical utility. Data privacy is achieved through the anonymisation of datasets before release. The risk of disclosure posed to the dataset should inform the level of anonymisation to be undertaken. As anonymisation achieves data privacy, it reduces the analytical utility of the dataset by introducing alterations to the original data values. Therefore, data curators require an appropriate estimate of the dataset's identity disclosure risk to inform the required anonymisation that balances privacy and utility. The disclosure risk varies from one geographical region to another due to varying enabling factors. This paper assesses the disclosure risk and the enabling factors in an environment lacking identified datasets in the public domain. This study used a quasi-experimental design in carrying out an empirical identity disclosure test, where respondents were given an anonymised dataset and were required to disclose the identity of any of the records. The findings were that background knowledge of the released datasets was the primary enabler in the absence of identified datasets. Respondents could only disclose records in the dataset they had familiarity with. However, the disclosure risk was within an acceptable threshold. Therefore, the study concluded that in an environment lacking identified datasets in the public domain, reasonable anonymisation could achieve a balance of privacy and utility in datasets. The findings justify private data release able to support data analytics and secondary data analyses in environments lacking identified datasets in the public domain.

APA CITATION

Muturi, P. N., Kahonge, A. M. & Chepken, C. K. (2022). Assessing Identity Disclosure Risk in the Absence of Identified Datasets in the Public Domain. *East African Journal of Information Technology*, *5*(1), 62-75. https://doi.org/10.37284/eajit.5.1.773

CHICAGO CITATION

Muturi, Peter N., Andrew M. Kahonge & Christopher Kipchumba Chepken. 2022. "Assessing Identity Disclosure Risk in the Absence of Identified Datasets in the Public Domain". *East African Journal of Information Technology* 5 (1), 62-75. https://doi.org/10.37284/eajit.5.1.773.

HARVARD CITATION

Muturi, P. N., Kahonge, A. M. & Chepken, C. K. (2022) "Assessing Identity Disclosure Risk in the Absence of Identified Datasets in the Public Domain", *East African Journal of Information Technology*, 5(1), pp. 62-75. doi: 10.37284/eajit.5.1.773.

IEEE CITATION

P. N. Muturi. A. M. Kahonge & C. K. Chepken "Assessing Identity Disclosure Risk in the Absence of Identified Datasets in the Public Domain", *EAJIT*, vol. 5, no. 1, pp. 62-75, Aug. 2022.

MLA CITATION

Muturi, Peter N., Andrew M. Kahonge & Christopher Kipchumba Chepken "Assessing Identity Disclosure Risk in the Absence of Identified Datasets in the Public Domain". *East African Journal of Education Studies*, Vol. 5, no. 1, Aug. 2022, pp. 62-75, doi:10.37284/eajit.5.1.773.

INTRODUCTION

The demand for sharing data for analytical purposes has continued to grow (Bandara et al., 2020; Templ et al., 2015). Data release for data analytics is an important aspect that is considered an enabler of unlocking the potential presented by the Big Data phenomena (Schroeck et al., 2012). Data analytics turns Big Data into a source of hindsight, insight, and foresight, which is crucial in an information-driven economy. The data release is also a fertile platform for breeding innovations (Nelson, 2015) and many more benefits (El Emam, Buckeridge, et al., 2011). The demand for secondary data analysis has also increased (Johnston, 2014; Wickham, 2019), raising the need for data release. The hurdle in realising data release by data curators is how to safeguard privacy and still retain analytical data utility.

Analytical data utility is viewed in terms of data accuracy and closeness to the original dataset (Domingo-ferrer et al., 2017; Reiter, 2015). The lesser the alterations are done to the dataset, the more the data utility is retained. However, high data accuracy raises the probability of data subjects being disclosed by an adversary (Antoniou et al., 2022). High data accuracy, i.e., data utility, threatens data privacy. Data privacy is achieved if the released data cannot be associated with high confidence with the individual subject. Data privacy is achieved through data anonymisation. There are many anonymisation techniques (Bandara et al., 2020; Domingo-ferrer et al., 2017), but they all involve some alterations of the original data, thereby affecting the data accuracy, hence reducing the data utility (Asikis & Pournaras, 2020; Domingo-ferrer et al., 2017; Erdélyi et al., 2018). Datasets with very high levels of privacy tend to have very low data utility (Asikis & Pournaras, 2020). Therefore, there is a need to strike a balance and tradeoff between data privacy and utility (Domingo-ferrer et al., 2017).

Anonymisation and Disclosure Risk

Releasing de-identified datasets thought to be anonymous has been shown to cause privacy breaches in some cases. Scenarios where adversaries have successfully re-identified record(s) from released datasets, assumed to be anonymous, have been documented (Li et al., 2007; Machanavajjhala et al., 2006; Sweeny, 2002). The disclosures may be a pointer to anonymisation that may have underestimated the risk of disclosure posed by the adversaries. On the other hand, overestimating the risk leads to more strict anonymisation, which erodes data utility (El Emam et al., 2009; Garfinkel, 2015) (El Emam & Hassan, 2016). Therefore, it is essential to assess and obtain the right estimate of the disclosure risk for a given dataset in a given geographical region to achieve the balance and tradeoff between privacy and utility.

The disclosure risk has been shown to be geographically dependent (Shlomo, 2009). Therefore, assessing the disclosure risk for a given region becomes a significant milestone towards achieving anonymisation that will preserve data privacy and retain analytical data utility. Some regions, especially the Western countries, have had a lot of research in the area of anonymisation and disclosure risk (Antoniou et al., 2022; Farzanehfar et al., 2021; Rocher et al., 2019; Santu et al., 2018; Sweeney et al., 2017, 2018; Xia et al., 2021; Yoo et al., 2018), but the same cannot be said of Kenya in East Africa. While many Western countries have had data protection laws in place (Quach et al., 2022), enabling them to release data in the public domain, Kenya enacted such a law in November 2019 (Parliament, 2019). Data release, more so in the public domain, was still relatively new at the time of this study. There was no study literature found on the level of the disclosure risk. Therefore, this study sought to assess the risk of identity disclosure in the region using an empirical quasi-experimental design.

REVIEWED LITERATURE

The need for data sharing for secondary analysis continues to be realised as the role of data analytics in the decision-making process becomes embraced. Sharing statistical microdata can enable numerous useful secondary analyses, which may be essential in supporting decision-making and policy formulation. But such data sharing must protect the privacy of the data subjects from whom data was collected. Most countries have a data protection legal framework requiring data to be anonymised before it is released to third parties or the general public. Examples of such legal frameworks include the European General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA) and Canadian Consumer Privacy Protection Act (CPPA) (Antoniou et al., 2022; Quach et al., 2022; Rocher et al., 2019). Along the same line, Kenya enacted the data protection act in November 2019 (Parliament, 2019). With the law now in place, its operationalisation is expected to create a conducive environment for data sharing. The focus of this research is on a category of recipients of such released datasets, who makes attempts, and intends to succeed in disclosing the data subjects. Such data recipients are referred to as adversaries or attackers (Kniola, 2017).

Some regions have very comprehensive laws and regulations on privacy preservation regarding data release. Some examples are Europe, Canada, California, and the USA in general (Khaled El Emam, Lucy Mosquera, 2020; Rocher et al., n.d.). The laws and regulations have enabled such regions to release various datasets that data analysts can use for secondary analyses and data analytics. However, some other regions have very weak or even lack framework on data release. As a result, such regions may tend to hoard datasets to avoid privacy breaches (Mitchell, 2012). Therefore, data privacy should be viewed in a regional context because what is considered private or could lead to privacy breaches in one region may not be in another region (Chakravorty et al., 2013). Indeed, research findings have shown different regions having varying levels of disclosure risk based on the same attributes (Fida Kamal Dankar et al., 2012; El Emam et al., 2009; El Emam, Buckeridge, et al., 2011).

Anonymisation and Privacy Concerns

Despite the legal framework requiring anonymisation of data before release and data curators following the guidelines, privacy concerns remain (Yao et al., 2016). Therefore, data curators must address privacy concerns before releasing data to third parties or the public. Indeed, research has shown that privacy breach concerns make many individuals not participate in data collection exercises (Cavoukian & Reed, 2013; Ramachandran et al., 2012), which may deprive vital research data. Therefore, data curators need to ensure the released datasets are anonymous to mitigate these concerns.

There are many anonymisation techniques that are used to protect privacy (Alfalayleh & Brankovic, 2015; Nissim et al., 2017; Reddy & Prakash, 2014). One approach involves attribute suppression. Suppression entails the removal of a field or a column in a dataset. It applies to unique identifiers and any other field consider not safe to be retained in the dataset. Another anonymisation approach is character masking, where some characters are replaced with a wild character such as an asterisk (*). Pseudonymisation is yet another anonymisation approach, where some or all characters are replaced with other values (Ribeiro & Nakamura, 2019). Then there is also generalisation, where data precision is reduced to achieve anonymisation. Swapping of values, where there is value rearrangement, is also used to anonymise. Finally, there is a perturbation approach, where some noise is introduced to the original data. All these techniques introduce different levels of alterations to the original data. They, therefore, affect the accuracy of the dataset by various degrees, hence reducing the data utility. The higher the privacy in a dataset, the lesser the utility and vice versa (Xia et al., 2021). That is why tradeoffs between the two must be made to balance the competing interests.

Despite all these anonymisation techniques aimed at having released data not linked to any specific individual, the risk is still there. Indeed, there are many instances where datasets were released to the public and resulted in a re-identification (Benitez & Malin, 2010; El Emam et al., 2009; El Emam, Buckeridge, et al., 2011; Ramachandran et al., 2012; Rocher et al., 2019; Sweeney, 2000), yet the data was thought to be anonymous. Such disclosures beg the question, when is the anonymisation adequate enough to make the dataset anonymous? The answer to the question lies in the data curator's understanding and being able to appropriately estimate the risk of disclosure posed to the dataset.

Disclosure Risk

The goal of the data curator is to release datasets in which there is no association of the released statistical data with the data subjects with high confidence. However, an adversary who receives the dataset seeks to learn more about the data subject as a result of interacting with the released dataset. The adversary achieves this by either 1) relying on background information they have combined with some of the data retained in the anonymous datasets released, or 2) linking externally available identified datasets with the released anonymous dataset, where the two have some common attributes. If an adversary gains information about a data subject that could not have been gained without interacting with the released datasets, then disclosure is said to have occurred. Therefore, disclosure risk is the probability of an adversary learning about an individual or group of individuals whose data is in the released anonymous dataset due to having accessed the datasets (Bambauer et al., 2014; El Emam & Alvarez, 2015; Lee & Clifton, 2012; Narayan, 2015).

Two broad categories of statistical disclosure are likely to happen: attribute disclosure and identity disclosure (Andreou et al., 2017). In attribute disclosure, also called attribution, an adversary can learn about a given attribute for a given group of participants without knowing their identity. For example, suppose participants share some common values for a given attribute, and the adversary can tell a participant belongs to that group. In that case, the adversary can learn about a participant without singling out any one of them. On the other hand, identity disclosure also referred to as re-identification, is where an adversary is able to map, with a high degree of confidence, a statistical data record to the individual to which it belongs (Andreou et al., 2017; Emam et al., 2020). It follows that every identity disclosure (re-identification) leads to an attribute disclosure, but the converse is not always the case. This study focused on identity disclosure (re-identification) since it poses more threat to the privacy of the data subjects.

Identity Disclosure Risk and Anonymisation

Three forms of identity disclosure risks are known to happen: prosecutor risk, marketer risk, and journalist risk (Xia et al., 2021). In prosecutor risk, the adversary has specific individual(s) they know are/are in the released datasets, and they target to re-identify them. In a marketer risk, the adversary has a certain group or calibre of individuals of interest they seek to re-identify, whereas, in a journalist risk, the adversary has no specific targets. They are only interested in matching any records from the released anonymised dataset with identified datasets to disclose the data subject's identity. In all these types of risk, the anonymised dataset plays a central role. The adversary seeks to use background knowledge or available identified datasets to uncover the identity of those in an anonymised dataset. The disclosure risk posed by the adversary to the datasets is therefore essential in determining the type and level of anonymisation to be used in anonymisation.

Identity disclosure is usually caused by underestimating the threat posed by the adversary, resulting in a not adequately anonymised dataset. On the other hand, overestimating the threat of disclosure will cause a high level of anonymisation, which is known to reduce the analytical utility of the released datasets (El Emam et al., 2009), (Xia et al., 2021). An appropriate estimation of the disclosure risk needs to be established in order to determine the appropriate levels of anonymisation that will be suitable for a given data release, to balance data privacy and utility,

Research Gap

There is a lot of research conducted on estimating the threat of identity disclosure in environments where identified datasets used for linkage/matching with anonymised datasets are available in the public domains (El Emam, Buckeridge, et al., 2011); (Benitez & Malin, 2010; Fida K. Dankar & El Emam, 2010; Domingo-Ferrer & Torra, 2003, 2004; El Emam et al., 2009; Ramachandran et al., 2012; Rocher et al., 2019; Scaiano et al., 2016; Simon et al., 2019). However, we have not come across research on estimating the risk in environments where identified datasets are not available in the public domain. The literature further showed that the disclosure risk is geographically dependent (Xia et al., 2021), with different regions having varying risk levels. This means each geographical region needs to do its disclosure risk assessment. In many of the studies cited, researchers used a theoretical approach to assess the disclosure risk. The theoretical approach assumes a worst-case scenario, where the adversary is assumed to have adequate tools and resources to carry out the re-identification. Assuming a worst-case scenario has resulted in an overrated risk, leading to anonymisation that sacrifices data utility (Xia et al., 2021).

Kenya is an example scenario where there aren’t many identified datasets in the public domain. We did not come across any research study from the region documenting the levels of disclosure risk and the attributes that enable or facilitate the risk. The fear of releasing data that will cause privacy breaches could be leading to data hoarding by data curators, hindering the realisation of the benefits of potential locked up in big data. The unquantified disclosure risk may result in failure to release data when there is no justifiable threat. Establishing the threat will help the data curators know the safe levels of anonymisation suitable for supporting data release for analytics (Xia et al., 2021).

This study presents an empirical approach to assessing the identity disclosure risk and identifying the factors facilitating the disclosure risk. The research employed an empirical quasi-experiment approach. The study was conducted in Kenya, using an educational dataset from sampled Kenyan University Students.

METHODOLOGY

The research adopted a quasi-experimental design and empirically conducted a test to assess the disclosure risk using respondents. The research had only one group of respondents who attempted to cause identity disclosure using the released anonymised dataset. The disclosure risk metric used in this study was the proportion of successful identity disclosure that occurred, which measured the vulnerability of records in the anonymised dataset (Benitez & Malin, 2010). The data collected was then quantitatively analysed.

Research Datasets

The study used educational data, specifically targeting data regarding University students in Kenya. University students are a well-recognisable group of individuals that would make it possible for respondents in the study to narrow the scope of data they need to search for as they attempt to disclose. Another consideration was that, except for a few students in their first year of study, all University students in Kenya are eighteen (18) years and above. At the age of eighteen years and above, an individual can give consent or decline to take part in a research study like this one. Therefore, this study excluded students who were in their first year of study to avoid those that may be underage. That ensured all participants were adults, capable of giving consent to be involved in the research.

Our efforts to get a secondary dataset from a data curator who regularly collected data about students were unsuccessful. We attributed the reluctance in data sharing to a legal framework that was not fully functional to guide data released to a third party at the time of the research in Kenya. Further, it was a statement of evidence on the lack of datasets in the public domain. We also realised there would have been hurdles and ethical concerns working with secondary data in that the data subject may not have given consent for their data to be used for this kind of research. We needed data with full disclosure so that it would be possible to verify whether any disclosure claimed to have taken place was indeed an accurate identity disclosure. That was the reason the study resulted in collecting primary data to be used for disclosure testing. Collecting primary data allowed seeking consent for data usage from the data subject.

Primary Identified Dataset

We purposively sampled five Kenyan universities and collected students' information. The five universities were sampled due to their proximity to each other, which was expected to provide background information to the would-be respondents. Three of the five were public universities, while two were private universities.

Researchers collected data after being granted permission following writing requests to the five universities. Researchers did a physical visit to each of the universities. With the assistance of various Heads of Departments, researchers identified groups of students who had classes on the day of the visits. The researchers randomly walked into lecture halls before lectures started and met with students. The purpose and the methodology of the research were explained to the students. Researchers then requested those willing to take part in the study to give their contact details. Those who gave their contacts were sent an online form to fill in their personally identifiable information, demographics, and academic-related information.

After researchers received the filled forms, they did data cleaning by removing duplications and those that had many missing values. The cleaned dataset became the identified dataset. The identified dataset had two hundred and sixty-six (266) records. The records were split into two datasets, each with one hundred and thirty-three (133) records and named Dataset 1 and Dataset 2.

Anonymised Datasets

The two datasets, Dataset 1 and Dataset 2 were anonymised using suppression and generalisation. Suppression was used in removing direct/explicit identifiers. Dataset 1 had a complete date of birth for each record, but in Dataset 2, we generalised the birth attribute by retaining only the year of birth. In Dataset 2, hobbies and home county attributes were suppressed by being dropped. The rest of the attributes in both datasets remained the same. Both datasets were given to respondents for the re-identification test exercise.

The attributes that were retained in the anonymised datasets were: Gender, Year/Date of Birth, Home County, Religion, University Enrolled in, University Campus enrolled in, Programme Enrolled, Faculty/School & Department, Course being Taken, Admission Year, current Year and Semester of study, Academic Progress Delay, Cause of Delay, Sponsorship, Applied Students' Loan, Given Student Loan, Accommodation when in Session, and Hobbies.

Sample Size

The respondents' sample size for the disclosure test was determined following Cohen's (Cohen, 1992) guideline at an effect size of 0.5, a statistical power of 90%, and a confidence level of 95%. Following this guideline, the minimum sample size was forty-four (44) respondents. However, the researchers surpassed that by having seventy-two (72) respondents for Dataset 1 and sixty-seven (67) for Dataset 2. Thus, the statistical values used gave this research findings excellent statistical significance.

The respondents who took part in the identity disclosure quasi-experiment (referred to as the test) were randomly sampled. The sample comprised students from the five universities from which the identified dataset was collected. Some of the sampled students had their data in the anonymised dataset. The sample also had non-student members from the general public.

Identity Disclosure Test

First, the respondents were given an orientation regarding the research and their role. Then, the respondents were given the anonymised datasets (Dataset 1 & Dataset 2), and they were asked to try to disclose the identity of any record(s) from the datasets. During the disclosure test, respondents were left free to use their means of re-identification. Respondents were told to look for any external datasets on their own that they may have needed/wanted to aid in the re-identification test.

The respondents were required to complete an online questionnaire in which they stated the record(s) they thought they had disclosed and provided the name(s) or any other identity of the individual(s) they thought they had disclosed. For every record claimed to have been re-identified by the respondent, reference to the identified dataset was made to verify whether, indeed, re-identification had occurred. The respondents' feedback on the disclosure test was compiled and analysed.

RESULTS

This section presents the analyses of the feedback obtained from the respondents who took part in the identity disclosure exercise. We first examine the demographics of the sampled respondents and then dive into the identity disclosures that took place. We then report on linkage datasets availability and how that influences disclosure risk. Finally, we report on the disclosure enablers relied on by the respondent in causing re-identification.

Respondents Demographics

Seventy-two (72) respondents completed the re-identification test using Dataset 1, students comprising 66.7% and 33.3% being general public (i.e. non-students). In terms of gender, 23.6% were female, while 76.4% were male respondents. The same respondents did the re-identification test using Dataset 2, but the number dropped to sixty-seven (67), comprising 64.2% of students and 35.8% general public. In terms of gender for Dataset 2, 22.4% were female, while 77.6% were male. The reduced number of respondents for Dataset 2 was due to the removal of some respondents who had left many blanks in their feedback to Dataset 2. Following the G-Power formula on sample size, at forty-four (44) respondents, the research would have 0.90 statistical power, an indicator of the confidence level of the results. Our sample surpassed that number.

Established Identity Disclosure Risk

The respondents reported the individuals they thought they had succeeded in re-identifying. The claim was verified if it was true or false. The outcome was either a true re-identification or no/false re-identification. The results of the re-identification test for Datasets 1 and 2 that took place are shown in *Table 1*.

Table 1: Re-identification Success Rate

Re-Identification		Dataset 1	Dataset 2
		Percent	**Percent**
Type	True	19.4	9.0
	False	80.6	91.0

Table 1 shows how respondents scored during the attempt to re-identify records from the anonymised Dataset 1 and Dataset 2. Successive re-identification was at 19.4% or 0.194 for Dataset 1, whereas Dataset 2 was 9% or 0.09. Further analysis to reveal the kind of re-identification that took place was done. The results are shown in *Table 2*.

Table 2: Re-identification Type Categorization

Re-identification		Dataset 1	Dataset 2
		Percent	**Percent**
Type	None	80.6	91.0
	Self	6.9	4.5
	Others	9.7	4.5
	Self & Others	2.8	

In *Table 2*, feedback on the re-identification type for each respondent who did the disclosure test was reported. It showed that 80.6% of the respondents did not re-identify any record from the anonymised Dataset 1. Those that only re-identified themselves from Dataset 1 were 6.9%. Those that were able to re-identify themselves and others were 2.8%. Thus, only 9.7% successfully re-identified others. The same test with Dataset 2 had 91% unable to re-identify any record, 4.5% re-identified themselves, and only 4.5% re-identified other people in the dataset.

Further analysis of the respondents' universities and the disclosed entities' universities was done. The analysis was to check if respondents could disclose records from universities other than those they were in or associated with. The results are shown in *Table 3*. The rows reflect the university a respondent was associated with, while the columns represent whether disclosure happened or not, and if it happened, the university the disclosed record came from.

Table 3: Respondents' University Association vs Re-identified Entity University

		Re-identified Entity University						**Total**
		None	**Uni.1**	**Uni.2**	**Uni.3**	**Uni.4**	**Uni.5**	
Respondent's University	None of the 5	2	0	0	0	0	0	2
	Uni.1	11	3	0	0	0	0	14
	Uni.2	5	0	0	0	0	0	5
	Uni.3	9	0	0	2	0	0	11
	Uni.4	3	0	0	0	1	0	4
	Uni.5	19	0	0	0	0	8	27
	More than One	9	0	0	0	0	0	9
Total		58	3	0	2	1	7	72

T*able 3* shows all the cases of re-identification that occurred involved respondents and entities from the same universities. For instance, respondents from University 1 (Uni.1) only re-identified entities (records) from University 1. The same thing for universities 3, 4, and 5. However, University 2 (Uni.2) respondents could not re-identify any records, including those from their university.

External Identified Datasets Linkage

Analysis of disclosure success and linkage to identified datasets used was done. The analysis was to determine which identified datasets were accessible to the respondents and utilised for linkage purposes during the re-identification exercise. *Table 4* shows the results. The columns show what the respondents cited as the external linkage datasets they used to assist them during the re-identification exercise, while the rows show whether the re-identification occurred or not.

Table 4: Re-Identification Success vs Linkage Datasets

		Linkage Datasets					**Total**
		None	**Class List/ Attendance**	**Google**	**Examination Marksheet**	**Student Records**	
Re-Identification Results	True	10	2	0	0	2	14
	False	52	0	1	1	4	58
Total		62	2	1	1	6	72

From *Table 4*, only 14.3% of all the respondents who successfully re-identified records used real external identified datasets for linkage purposes. However, the external identified datasets used were not in the public domain. Thus, at 85.7%, most disclosure cases took place without respondents using any identified datasets for linkage purposes.

Identity Disclosure Enablers

Respondents had been asked to state the enablers they used to aid the disclosure exercise. The results are shown in *Table 5.*

Table 5: Re-Identification Enablers and Usage

Enablers		Gender	DoB	Religion	County	Familiarity
Usage (%)	Yes	12.5	19.4	4.2	12.5	34.7
	No	87.5	80.6	95.8	87.5	65.3
	Total	100.0	100.0	100.0	100.0	100.0

Table 5 shows that the most used enabler (34.7%) was familiar with the provided datasets during the disclosure exercise. Familiarity in this context represented the background knowledge the respondent possessed. The date of birth (DoB) attribute was second, with 19.4% of respondents using it. Gender and County tied at 12.5%, and religion at only 4.2% of respondents using it during the disclosure exercise.

DISCUSSION

The study intended to empirically establish identity disclosure risk levels in a region where identified datasets are not readily available in the public domain for linkage purposes. As of the time of the study in Kenya, there was not much data released to the public that could support data analytics or secondary analysis. The findings established the disclosure risk levels as well as identified the enablers of re-identification in the region. The results are essential to data curators and policymakers concerning data privacy and data release/sharing.

Anonymisation and Balancing Privacy & Utility

There had been a general perception that institutions and agencies that collect and store data in Kenya could not share that data due to the risk of data subjects being re-identified. However, from the results of this study, a simple form of de-identification and data generalisation proved to be adequate for anonymisation. *Table 1* shows the overall disclosure risk was 19.4%, as reflected by the successful identity disclosure rate. The overall rate is inclusive of self-disclosure. Since self-disclosure does not result in a privacy breach, it should be left out, lowering the disclosure risk to 12.5% in Dataset 1.

When the same respondents used Dataset 2, which had the year of birth instead of the complete date of birth, and two more attributes (County and Hobbies) removed, the overall disclosure risk dropped to 9%. In this case, only 4.5% of the disclosure risk amounted to privacy breaches since they were not self-disclosure. The actual disclosure risk was, therefore, 4.5%. The generalisation of the birth attribute from the date of birth to the year of birth and removal of County and hobbies attributes significantly changed the identity disclosure success rate. It dropped from 12.5% to 4.5%.

From the baseline given in a study done (El Emam, Jonker, et al., 2011; Santu et al., 2018), the re-identification success rate was found to be 0.262 (or 26.2%) for all studies and 0.338 (or 33.8%) for health data. Further, the European Medicine Agency (EMA) and Health Canada have set an acceptable re-identification risk threshold at 0.09 (or 9%) (Branson et al., 2020). We, therefore, find the level of disclosure risk established for this study in Kenya, being 12.5% for Dataset 1 & 4.5% for Dataset 2, suggests the disclosure risk can be maintained within the acceptable re-identification threshold of 9%. Dataset 2 achieved the threshold quite well. Therefore, the disclosure risk that was in existence in the region posed no alarming danger to privacy breaches. Therefore, the findings should encourage the release of private datasets to support secondary analysis and data analytics, given the current environment in the region.

From the results in *Table 3*, respondents could only disclose records from the same universities they were in or were associated with. By association, it meant either working in that university or being an alumnus who had left the institution less than four years ago at the time of the research. Therefore, we concluded that if anonymised datasets are released to data users in environments with no identified datasets and data users' don't have background knowledge of the released datasets, the disclosure risk would be almost zero.

Absence of Linkage Datasets Influencing Re-identification

Table 4 exemplifies the lack of identified datasets for the respondents at the time of the disclosure exercise. A majority (85.7%) of the respondents who disclosed identities of some records did not refer to any identified datasets. Those that used external datasets for linkage used private data records (i.e. Student records) that are not accessible by the general public. Some respondents mentioned the Class List and Examination Marksheet in *Table 4* as identified datasets they used to aid disclosure. But those documents do not contain students' attributes that can cause disclosure. Class lists and Marksheets could only assist in confirming an individual's name that one had been able to re-identify. Student Records could undoubtedly be used for attribute matching leading to re-identification, but this is a resource whose access and use are guarded by the Universities and is not accessible even to most workers in the University setting. Among the respondents sampled, some were employees of some of the Universities that manage students' records. They used student records for linkage, and some were able to make successful re-identification.

The fact that 85.7% of disclosure happened without reference to any identified datasets implied that the disclosure was the prosecutorial type. In a prosecutorial disclosure attack, the adversary has background information about specifically targeted individuals he/she knows their data is in the dataset (Assuncao et al., 2016; El Emam et al., 2009; Emam, 2013; Garfinkel, 2015). Thus, the adversary relies on the knowledge of his/her target to locate them in the anonymised dataset. In prosecutor disclosure attacks, background information is the primary disclosure enabler. *Table 4* showed most respondents didn't rely on identified datasets to cause disclosure. Further, *Table 5* re-affirmed that background knowledge was the leading enabler. Familiarity with the released datasets was pointed out by many as their enabler during the disclosure exercise. That explains why respondents were only able to re-identify records from their universities.

From the findings presented, as long as no identified datasets are being released to the public that may be used for linkage, releasing anonymised datasets doesn't pose an unreasonable danger of privacy breach. With reasonable anonymisation, we have illustrated one can achieve acceptable levels of identity disclosure risk that balance privacy and utility. Indeed, the risk is almost zero if the data release targets specific data users known not to have background information on the released datasets. The fact that none of the respondents could identify any record from a university they were not in (*Table 3*) shows that data releases targeting users without background familiarity are unlikely to cause a privacy breach.

CONCLUSION

It was evident there were no identified datasets to be used for linkage purposes during the disclosure exercise available in the public domain in Kenya, the region where the research was carried out. The findings have demonstrated that in the absence of identified datasets in the public domain, suppression and data generalisation of a dataset is sufficient in achieving anonymisation. Moreover, the anonymisation did not alter the original datasets very much. Hence the data utility of the dataset was not sacrificed. Therefore, the balance of privacy and utility was achieved. The results further established that the adversary's familiarity with the released dataset was the leading enabler to identity disclosure in the absence of identification datasets.

There is no danger of privacy breach in environments that don't have identified datasets in the public domain, and it's known that the data users don't have background information (familiarity) on the anonymised dataset released. Therefore, releasing anonymised datasets in such regions doesn't put individuals whose data is released at unreasonable risk.

ACKNOWLEDGEMENTS

We wish to express our gratitude to the students and members of the public who participated in our research. We also thank the management of the five universities for allowing the research to be conducted in their respective institutions. Equally, we are grateful to the National Commission for Science, Technology and Innovation (NACOSTI) for granting the research permit. Lastly but not least, our appreciation to Dr. Bonface N. Ireri for his editorial and moral support.

REFERENCES

Alfalayleh, M., & Brankovic, L. (2015). Quantifying privacy: A novel entropy-based measure of disclosure risk. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8986*, 24–36. https://doi.org/10.1007/978-3-319-19315-1_3

Andreou, A., Goga, O., & Loiseau, P. (2017). Identity vs. Attribute disclosure risks for users with multiple social profiles. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 163–170. https://doi.org/10.1145/3110025.3110046

Antoniou, A., Dossena, G., Macmillan, J., Hamblin, S., Clifton, D., & Petrone, P. (2022). Assessing the risk of re-identification arising from an attack on anonymised data. In *arXiv:2203.16921*. https://arxiv.org/ftp/arxiv/papers/2203/2203.16921.pdf

Asikis, T., & Pournaras, E. (2020). Optimisation of privacy-utility tradeoffs under informational self-determination. *Future Generation Computer Systems*, *109*, 488–499. https://doi.org/10.1016/j.future.2018.07.018

Assuncao, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., Buyya, R., Avery, A. A., Cheek, K., Bailey, D., Bailey, S., Bâra, A., Lungu, I., Bari, A., Chaouchi, M., Jung, T., For, A., Berk, Bordawekar, R., Blainey, B., Apte, C., … Zementis. (2016). Why Your Next Data Warehouse should be in the Cloud. *Going Pro in Data Science*, *25186*(June), 1–7. https://doi.org/10.1126/science.Liquids

Bambauer, J., Muralidhar, K., & Sarathy, R. (2014). Fool' s Gold : An Illustrated Critique of Differential Privacy. *Vanderbilt Journal of Entertainment & Technology Law*, *16*(4), 701–755. https://scholarship.law.vanderbilt.edu/jetlaw/vol16/iss4/1/

Bandara, P. L. M. K., Bandara, H. D., & Fernando, S. (2020). Evaluation of Re-identification Risks in Data Anonymization Techniques Based on Population Uniqueness. *Proceedings of ICITR 2020 - 5th International Conference on Information Technology Research: Towards the New Digital Enlightenment*. https://doi.org/10.1109/ICITR51448.2020.9310884

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, *17*(2), 169–177. https://doi.org/10.1136/jamia.2009.000026

Branson, J., Good, N., Chen, J. W., Monge, W., Probst, C., & El Emam, K. (2020). Evaluating the re-identification risk of a clinical study report anonymised under EMA Policy 0070 and Health Canada Regulations. *Trials*, *21*(1), 1–9. https://doi.org/10.1186/s13063-020-4120-y

Cavoukian, A., & Reed, D. (2013). Big Privacy: Bridging Big Data and the Personal Data Ecosystem Through Privacy by Design. In *Information and Privacy Commissioner of Ontario, Canada* (Issue December). www.ipc.on.ca/images/Resources/pbd-big_privacy.pdf

Chakravorty, A., Wlodarczyk, T., & Rong, C. (2013). Privacy preserving data analytics for smart homes. *Proceedings - IEEE CS Security and Privacy Workshops, SPW 2013*, 23–27. https://doi.org/10.1109/SPW.2013.22

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1).

Dankar, Fida K., & El Emam, K. (2010). A Methods of Evaluating Marketer Re-Identification Risk. *ACM International Conference Proceeding Series*.

Dankar, Fida Kamal, El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC Med Inform Decis Mak*, *12*(September 2009), 66. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583146/?tool=pubmed%5Cnhttp://dx.doi.org/10.1186/1472-6947-12-66

Domingo-ferrer, J., Ricci, S., & Soria-coma, J. (2017). Empirical Comparison of Anonymisation Methods Regarding Their Risk-Utility Tradeoff. *International Conference on Modeling Decisions for Artificial Intelligence*, 1–15. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2017/3_utility_risk.pdf

Domingo-Ferrer, J., & Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, *13*(4), 343–354. https://doi.org/10.1023/A:1025666923033

Domingo-Ferrer, J., & Torra, V. (2004). Disclosure risk assessment in statistical data protection. *Journal of Computational and Applied Mathematics*, *164–165*, 285–293. https://doi.org/10.1016/S0377-0427(03)00643-5

El Emam, K., & Alvarez, C. (2015). A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymisation techniques. *International Data Privacy Law*, *5*(1), 73–87. https://doi.org/10.1093/idpl/ipu033

El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., & Verma, A. (2011). The re-identification risk of Canadians from longitudinal demographics. *BMC Medical Informatics and Decision Making*, *11*(1). https://doi.org/10.1186/1472-6947-11-46

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., & Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *Canadian Journal of Hospital Pharmacy*, *62*(4), 307–319. https://doi.org/10.4212/cjhp.v62i4.812

El Emam, K., & Hassan, W. (2016). *The De-identification Maturity*. https://www.himss.org/privacy-analytics-de-identification-maturity-model

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0028071

Emam, K. El, Mosquera, L., & Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *Journal of Medical Internet Research*, *22*(11), 1–14. https://doi.org/10.2196/23139

Emam, K. (2013). Measuring the Probability of Re--Identification. In *Guide to the De-Identification of Personal Health Information* (pp. 177–196). https://doi.org/10.1201/b14764-20

Erdélyi, Á., Winkler, T., & Rinner, B. (2018). Privacy protection vs. utility in visual data: An objective evaluation framework. *Multimedia Tools and Applications*, *77*(2), 2285–2312. https://doi.org/10.1007/s11042-016-4337-7

Farzanehfar, A., Houssiau, F., & de Montjoye, Y. A. (2021). The risk of re-identification remains high even in country-scale location datasets. *Patterns*, *2*(3), 100204. https://doi.org/10.1016/j.patter.2021.100204

Garfinkel, S. L. (2015). NISTIR 8053 De - Identification of Personal Information NISTIR 8053 De - Identification of Personal Information. In *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.IR.8053

Johnston, M. P. (2014). Secondary Data Analysis : A Method of which the Time Has Come. *Qualitatve and Quantative Methods in Libraryes (QQML)*, *3*, 619–626.

Khaled El Emam, Lucy Mosquera, R. H. (2020). *Practical Synthetic Data Generation*. O'Reilly Media, Inc. https://www.oreilly.com/library/view/practical-synthetic-data/9781492072737/

Kniola, L. (2017). Plausible Adversaries in Re-Identification Risk Assessment. *Phuse*, *Paper DH09*, 1– 10. https://www.lexjansen.com/phuse/2017/dh/DH09.pdf

Lee, J., & Clifton, C. (2012). Differential identifiability. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’12*, 1041. https://doi.org/10.1145/2339530.2339695

Li, N., Li, T., & Venkatasubramania, S. (2007). t -Closeness : Privacy Beyond k -Anonymity and -Diversity. *IEEE 23rd International Conference*, *3*, 106–115. https://doi.org/10.1109/ICDE.2007.367856

Machanavajjhala, A., Kifer, D., Gehrhe, J., & VENKITASUBRAMANIAM, M. (2006). L-Diversity : Privacy Beyond k -Anonymity. *Proceedings of the 22nd International Conference on Data Engineering*, 1–36.

Mitchell, A. (2012). From data hoarding to data sharing. *Journal of Direct, Data and Digital Marketing Practice*, *13*(4), 325–334. https://doi.org/10.1057/dddmp.2012.3

Narayan, A. (2015). *Distributed Differential Privacy and Applications*.

Nelson, G. S. (2015). Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. *SAS® Global Forum 2015*, *April 2015*, 23. http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf

Nissim, K., Steinke, T., Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., O ’brien, D. R., & Vadhan, S. (2017). *Differential Privacy: A Primer for a Non-technical Audience * (Preliminary version)*. *1237235*.

Parliament, K. (2019). The Data Protection Act. In *National Council for Law Reporting*. https://doi.org/10.1088/0031-9112/37/4/026

Quach, S., Thaichon, P., Martin, K. D., Weaven, S., & Palmatier, R. W. (2022). Digital technologies: tensions in privacy and data. *Journal of the Academy of Marketing Science*. https://doi.org/10.1007/s11747-022-00845-y

Ramachandran, A., Singh, L., Porter, E., & Nagle, F. (2012). Exploring re-identification risks in public domains. *2012 10th Annual International Conference on Privacy, Security and Trust, PST 2012*, 35– 42. https://doi.org/10.1109/PST.2012.6297917

Reddy, S., & Prakash, O. (2014). UTILITY-PRIVACY TRADEOFF IN DATABASES : AN INFORMATION THEORETIC APPROACH. *International Journal of Engineering & Science Research*, *4*(10), 608–612.

Reiter, J. P. (2015). Estimating Risks of Identification Disclosure in Microdata. *Journal of the American Statistical Association*, *100*(472), 1103–1112.

Ribeiro, S. L., & Nakamura, E. T. (2019). Privacy Protection with Pseudonymization and Anonymization in a Health IoT System: Results from OCARIoT. *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*, 904–908. https://doi.org/10.1109/BIBE.2019.00169

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10933-3

Rocher, L., Hendrickx, J. M., & Montjoye, Y. De. (n.d.). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications, 2019*. https://doi.org/10.1038/s41467-019-10933-3

Santu, S. K. K., Bindschadler, V., Zhai, C., & Gunter, C. A. (2018). NRF : A Naive Re-identification Framework. *Proceedings of the 2018 Workshop on Privacy in the Electronic Society (WPES’18)*, 121–132.

Scaiano, M., Middleton, G., Arbuckle, L., Kolhatkar, V., Peyton, L., Dowling, M., Gipson, D. S., & El Emam, K. (2016). A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of Biomedical Informatics*, *63*, 174–183. https://doi.org/10.1016/j.jbi.2016.07.015

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of big data. *IBM Global Business Services Saïd Business School at the University of Oxford*, 1–20. http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf

Shlomo, N. (2009). Releasing microdata: disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality*, *1*, 229–240. http://eprints.soton.ac.uk/65423/

Simon, G. E., Shortreed, S. M., Coley, R. Y., Penfold, R. B., Rossom, R. C., Waitzfelder, B. E., Sanchez, K., & Lynch, F. L. (2019). Assessing and Minimising Re-identification Risk in Research Data Derived from Health Care Records. *EGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *7*(1), 1–9. https://doi.org/10.5334/egems.270

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000*, 1–34. http://dataprivacylab.org/projects/identifiability/paper1.pdf

Sweeney, L., Loewenfeldt, M. V, & Perry, M. (2018). Saying it’s anonymous doesn’t make it so: Re-identifications of {\textquotedblleft} anonymized{\textquotedblright} law school data. *Technology Science*, *2018111301*. https://techscience.org/a/2018111301

Sweeney, L., Yoo, J. S., Perovich, L., Boronow, K. E., Brown, P., & Brody, J. G. (2017). Re-identification risks in HIPAA Safe Harbor data: a study of data from one environmental health study. *Technol Sci*, *2017:20170*.

Sweeny, L. (2002). k- ANONYMITY: A MODEL FOR PROTECTING PRIVACY 1. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, *67*(4). https://doi.org/10.18637/jss.v067.i04

Wickham, R. J. (2019). Secondary Analysis Research. *Journal of the Advanced Practitioner in Oncology*, *10*(4), 395–400. https://doi.org/10.6004/jadpro.2019.10.4.7

Xia, W., Liu, Y., Wan, Z., Vorobeychik, Y., Kantacioglu, M., Nyemba, S., Clayton, E. W., & Malin, B. A. (2021). Enabling realistic health data re-identification risk assessment through adversarial modeling. *Journal of the American Medical Informatics Association : JAMIA*, *28*(4), 744–752. https://doi.org/10.1093/jamia/ocaa327

Yao, X., Zhou, X., & Ma, J. (2016). Differential Privacy of Big Data: An Overview. *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, *9*(2), 7–12. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.9

Yoo, J. S., Thaler, A., Sweeney, L., & Zang, J. (2018). Risks to Patient Privacy : A Re-identification of Patients in Maine and Vermont Statewide Hospital Data. *Technology Science*, 1–62. https://techscience.org/a/2018100901/