The African Journal of Information Systems

Volume 15 | Issue 4 Article 1

November 2023

Modeling Identity Disclosure Risk Estimation Using Kenyan Situation

Peter N. Muturi
University of Nairobi, Kenya & Multimedia University of Kenya, pmuturi@mmu.ac.ke

Andrew M. Kahonge
University of Nairobi, Kenya, andrew.mwaura@uonbi.ac.ke

Christopher K. Chepken
University of Nairobi, Kenya, chepken@uonbi.ac.ke

Evans K. Miriti
University of Nairobi, Kenya, eamiriti@uonbi.ac.ke

Recommended Citation

Muturi, Peter N.; Kahonge, Andrew M.; Chepken, Christopher K.; and Miriti, Evans K. (2023) "Modeling Identity Disclosure Risk Estimation Using Kenyan Situation," *The African Journal of Information Systems*: Vol. 15: Iss. 4, Article 1.
Available at: https://digitalcommons.kennesaw.edu/ajis/vol15/iss4/1

The African Journal of Information Systems

Modeling Identity Disclosure Risk Estimation Using Kenyan Situation

Research Paper

Volume 15, Issue 4, November 2023, ISSN 1936-0282

Peter N. Muturi
Department of Computing and Informatics
University of Nairobi, Kenya
pmuturi@mmu.ac.ke

Andrew M. Kahonge
Department of Computing and Informatics
University of Nairobi, Kenya
andrew.mwaura@uonbi.ac.ke

Christopher K. Chepken
Department of Computing and Informatics
University of Nairobi, Kenya
chepken@uonbi.ac.ke

Evans K. Miriti
Department of Computing and Informatics
University of Nairobi, Kenya
eamiriti@uonbi.ac.ke

(*Received June 2022, accepted August 2023*)

ABSTRACT

Identity disclosure risk is an essential consideration in data anonymization aimed at preserving privacy and utility. The risk is regionally dependent. Therefore, there is a need for a regional empirical approach in addition to a theoretical approach in modeling disclosure risk estimation. Reviewed literature pointed to three influencers of the risk. However, we did not find literature on the combined effects of the three influencers and their predictive power. To fill the gap, this study modeled the risk estimation predicated on the combined effect of the three predictors using the Kenyan situation. The study validated the model by conducting an actual re-identification quasi-experiment. The adversary's analytical competence, distinguishing power of the anonymized datasets, and linkage mapping of the identified datasets are presented as the predictors of the risk estimation. For each predictor, manifest variables are presented. Our presented model extends previous models and is capable of producing a realistic risk estimation.

Keywords

Data analytics, anonymized data, de-identification, data privacy, data release, data sharing.

INTRODUCTION

As advancements in data collection, storage, and processing technologies are being realized, there is a rise in privacy concerns from both data subjects (individuals from whom data is collected) and the data custodians (Zlatolas et al., 2022). The need to monetize the data, which has led to data being released to third parties and even to the general public, has made the concerns even greater (Quach et al., 2022). Those concerns have necessitated governments and organizations to develop data protection regulations and frameworks, though that has not eliminated the concerns. The regulations and frameworks usually require datasets to be de-identified before release. De-identification ensures all personally identifiable information is removed from the dataset before the dataset is released. However, de-identification only reduces the chances of dataset recipients disclosing the identity of individuals in the dataset. The level of de-identification and concealing necessary to make a dataset anonymous should correspond to the risk of identity disclosure posed (Orooji & Knapp, 2018). Therefore, the prevailing risk of identity disclosure influences the anonymization process required to prepare a dataset for release. Furthermore, the identity disclosure risk is influenced by situation-dependent factors (Xia et al., 2021). That means the risk level will vary from one situation to another. Further, reviewed literature has shown that the identity disclosure risk is regionally dependent. Many studies in developed countries have been undertaken, and identity disclosure risks have been established (Antoniou et al., 2022; Bandara et al., 2020; El Emam et al., 2020; Farzanehfar et al., 2021; Ribeiro & Nakamura, 2019; Rocher et al., 2019; Simon et al., 2019; Xia et al., 2021). However, our study did not find any similar study for Kenya or its neighbors, establishing that as a gap that needs filling. This study sought to address this by identifying the factors influencing identity disclosure risk and developing a model for estimating the risk. The model was validated using empirical data from an identity disclosure experiment.

The remainder of this paper presenting the study is structured as follows: the literature review that led to the formulation of a conceptual model, study methodology, results, discussion and conclusion.

REVIEWED LITERATURE

With the convergence of technologies, large volumes of data are rapidly generated and stored, making the big data phenomenon a reality. But the phenomenon's potential remains untapped unless there are means of extracting useful information from that data. Efforts to leverage the big data phenomenon have led to data analytics being applied widely as the information-driven economy becomes predominant in many parts of the world (McKinsey Analytics, 2016).

Data analytics often requires analysts to have access to multiple datasets so that they are able to establish patterns and trends, leading to hindsight, insight, and foresight (El Emam et al., 2011; Nelson, 2015; Schroeck et al., 2012). Hence, as data analytics takes root, the need for data sharing has become critical. Data sharing entails the data custodian releasing datasets to either third parties or the general public for various uses, including secondary analysis. Secondary analysis is where a party other than the one who originally collected data for primary analysis is given access to the dataset for further analysis. This secondary analysis is usually different from the primary analysis. Indeed, secondary data analysis is becoming common practice and data sharing and data release should be promoted to leverage data analytics and secondary data analysis (Branson et al., 2020; Mello et al., 2018).

Data privacy and analytical utility are fundamental aspects of data sharing and release (El Emam & Hassan, 2016). The released dataset should be anonymous to the extent that it does not cause a privacy breach but remains useful for analytics purposes. If privacy preservation were the only concern, it would easily be achieved (Hsu et al., 2014). For example, frequency tables, contingency tables, and histograms are highly summarized, maintaining very high privacy but retaining little analytical utility. However, the act of balancing privacy and utility is what makes privacy preservation a non-trivial matter.

Privacy concerns have led many governments to develop privacy and data protection regulatory frameworks. Examples include Europe's General Data Protection Regulation, the USA Health Information Portability and Accountability Act, Safe Harbor, and California's Consumer Privacy Act (Rocher et al., 2019). All are aimed at protecting data subjects' confidentiality. Kenya enacted a data protection law in 2019 (Republic of Kenya, 2019), giving effect to articles 31(c) and (d) of the

Constitution of Kenya 2010 (Constitution of Kenya, 2010). But privacy and data protection require a holistic approach, including legal, administrative, and technical safeguards and the aforementioned legal safeguards require datasets to be anonymized before they are released. Such releases are considered to be private data releases. Anonymization involves removing explicit identifiers such as name, e-mail address, etc., that uniquely identify data subjects, but retaining some quasi-identifiers such as gender and year of birth, which, on their own, do not uniquely identify an individual in the dataset. However, reviewed literature has shown that some combinations of the quasi-identifiers have led to privacy breaches, with individuals in anonymous datasets being re-identified (Li et al., 2007; Machanavajjhala et al., 2007; Mello et al., 2018; Sweeney, 2002). That means having an anonymous dataset does not entirely eliminate disclosure risk.

Statistical Disclosure Risk

Disclosure risk refers to the probability of an adversary or attacker gaining information about a particular data subject by interacting with released anonymous datasets. An adversary or attacker in this context is any dataset recipient, intended or unintended, with motives to attempt, and aims to succeed in causing disclosure of a data subject (Kniola, 2017). Data subject refers to the individual from whom data was collected, in this case, represented by a record in the datasets. Disclosure risk is categorized into attribute disclosure and identity disclosure (also called re-identification). In attribute disclosure, an adversary successfully associates a particular attribute in the anonymous datasets to a specific population unit. This way, though unable to identify a specific data subject in the anonymous dataset, the adversary can gain attribute knowledge about some group of data subjects by interacting with the dataset. For example, suppose an adversary knows a given individual is among the records under consideration. In that case, the adversary may learn about an attribute of that individual they did not know about. Identity disclosure, also called re-identification, is when a given record in the released anonymous dataset is successfully associated with a specific data subject. In other words, the identity of an individual, who was initially unknown, is revealed. Our study focused on identity disclosure, i.e., the re-identification risk.

Most reviewed studies on disclosure risk assessment are for a specific geographical region. In this context, a region is an area of jurisdiction under a certain governance structure, legal framework, and policies. Examples of a region include a state, country, or group of countries forming an economic block. For example, some studies on disclosure risk assessment were based in Canada (Dankar et al., 2012; El Emam et al., 2009, 2010; El Emam, 2013; El Emam et al., 2020), others were based in the USA (Xia et al., 2021; Rocher et al., 2019; Simon et al., 2019), and one was based in Singapore (Personal Data Protection Commission Singapore, 2018). Regional disclosure risk assessment is logical because a major factor that influences disclosure risk depends on the attributes left in the anonymous datasets, and these could vary from region to region. For example, geographic information is one of the quasi-identifier attributes relied on in re-identification. The geographic information given will vary from one region to another and could have different disclosure abilities. For instance, the USA zip code, combined with a year of birth and gender, can re-identify up to 63% to 87% of the population in the USA (Benitez & Malin, 2010). The zip code provides details on the residential area of a given individual in the USA. Other regions, like Kenya, do not have such well-defined geographic information on residential areas. Instead, there are administrative blocks (counties, sub-counties, locations, and villages) that have many individuals residing there. Therefore, Kenyan geographic information may have much less disclosure information than the USA zip code.

Geographic information is just one example of how quasi-identifiers retained in the anonymized datasets that are shared or released could have different disclosure abilities. Indeed, Rocher et al. (2019) found fifteen demographic attributes that could be used to correctly re-identify 99.98% of Americans. However, different regions may have similar or completely different attributes each with varying disclosure strength.

Another influence of disclosure risk is the availability of identified datasets that adversaries can easily link with the released anonymized datasets to disclose the identity of data subjects. Some regions, such as the USA, make available voters' register information and this has been relied on by attackers for re-identification (Barth-Jones, 2012; Benitez & Malin, 2010; El Emam, 2006; El Emam et al., 2011; El Emam et al., 2020; Simon et al., 2020; Sweeney et al., 2017; Xia et al., 2021). However, in other regions, such datasets are not made public (Muturi et al., 2022). Further, some regions have well-structured legal frameworks and policies to guide anonymization of datasets before sharing or release. A good example is the EU's General Data Protection Regulation (Jayasinghe et al., 2019; Ribeiro & Nakamura, 2019; Rocher et al., 2019; Xia et al., 2021) and the USA Health Information Portability and Accountability Act (Dankar et al., 2012; Rocher et al., 2019; Santu et al., 2018). However, other regions do not have such frameworks.

We, therefore, claim that different regions will have varying disclosure risk levels and that the risk is regionally dependent. Indeed, it has been said to be situationally dependent (Xia et al., 2021). We base our claim on the fact that different regions will have different quasi-identifier attributes retained in the released datasets, and those attributes have different disclosure abilities. Further, different regions will have different laws and policies on data anonymization and release, which will influence the disclosure risk. Data custodians or curators in a region should, therefore, seek to establish the prevailing disclosure risk and identify attributes with high disclosure abilities.

Since a risk management strategy requires understanding the threat posed so as to know how to avert it, each region needs to establish a realistic estimate of its prevailing disclosure risk. There are many studies in the USA, Canada, and other developed countries relating to disclosure risk (El Emam et al., 2011; El Emam et al., 2020; Patel & Jethava, 2018; Rocher et al., 2019; Xia et al., 2021), but the same cannot be said of African countries. For example, our study did not find any literature on modeling disclosure risk in Kenya or her neighboring East African countries. Therefore, there is a need for an empirical approach to modeling the disclosure risk for Kenya in order to get realistic disclosure estimates. Relying on a theoretical approach, which adopts a worst-case scenario during disclosure risk estimation, tends to provide higher disclosure risk estimates than there may really be (Xia et al., 2021). Overestimates of disclosure risk produced by the theoretical approach can lead to an unnecessarily stringent level of anonymization, meaning higher privacy levels at the expense of analytical utility. Hence, an empirical approach is needed to estimate a realistic or actual risk to guide anonymization aimed at achieving a balance between privacy and utility. Keeping data private means that the risk of disclosure is maintained at a level that is lower than or equal to a given acceptable threshold, currently set at 0.09 and applied even in the medical research fields (Branson et al., 2020).

To achieve a private data release that preserves both privacy and analytical utility, the data custodian or curator needs to understand the prevailing disclosure risk. Underestimating that risk will lead to anonymization that could easily cause privacy breaches. On the other hand, overestimating the risk will lead to too much suppression of the anonymous datasets, reducing the dataset's analytical utility (Xia et al., 2021). The need to balance privacy and utility underscores the necessity of a realistic estimate of the disclosure risk.

Reviewed literature has identified several factors that influence identity disclosure risk. They revolve around three players: the adversary, the released anonymous datasets, and the auxiliary identified datasets used for LM with the anonymous datasets leading to disclosure (Dankar et al., 2012; Elliot, 2000; Mello et al., 2018; Rocher et al., 2019; Skinner & Elliot, 2002; Sweeney, 2000; Xia et al., 2021). However, our study did not find any model that sought to study the combined effects of the factors influencing disclosure risk and their predictive strengths. Therefore, there is a need to develop a framework for modeling disclosure risk prevailing in a given region (country or state) to establish realistic disclosure risk estimates.

The Study Conceptual Model

Statistical disclosure risk is the probability that a record that was anonymous (de-identified) in the released dataset would be disclosed (re-identified) by an adversary due to interaction with the released dataset. Our conceptual model is, therefore, based on probability theory. Probability theory uses random variables and probability distributions to mathematically evaluate the probability associated with a random phenomenon. A random variable is one whose value is unknown a priori, and a probability distribution is a mathematical function describing the probability of different possible values of a variable. A random phenomenon is a situation in which the possible outcomes are known but not the specific ones that will happen. Probability theory uses formal concepts to describe the probability of occurrence of a given outcome. The random phenomenon and the random variables influencing the phenomenon need to be established for probability theory to be applied. The probability distribution then gives the relationship between the random variables and the random phenomenon.

Probability theory can be approached either theoretically or experimentally. Theoretical probability is determined through logical reasoning without conducting experiments. Experimental probability, on the other hand, is determined as a result of data obtained from repeated experiments. Our research adopted experimental probability to get empirical data based on actual disclosure attempts.

The reviewed literature on assessing or estimating disclosure risk has applied probability theory, though not stating it explicitly, mostly using theoretical probability (Bandara et al., 2020; Bethlehem et al., 1990; Domingo-Ferrer & Torra, 2003, 2004; Duncan & Lambert, 1989; Manrique-Vallier & Reiter, 2012; Paass, 1988; Shlomo, 2010; Shlomo & Skinner, 2022; Skinner & Elliot, 2002; Xia et al., 2021). From the literature reviewed, there are three players from whom random variables are drawn. These are the adversary (Wan et al., 2015; Xia et al., 2021), released dataset (Kounine & Bezzi, 2008; Lubarsky, 2017; Winkler, 2005), and auxiliary identified datasets (Barth-Jones, 2012; Benitez & Malin, 2010; Machanavajjhala, 2007 ; Sweeney, 2002; Xia et al., 2021). The random phenomenon is the risk of disclosure, mainly identity disclosure (re-identification). Some researchers have modeled the disclosure risk based on unique attributes of both the sampled dataset and the population and the linkage matching key attributes between the released dataset and the auxiliary datasets. Others have modeled disclosure risk based on the adversary's capabilities and resources (Xia et al., 2021). The literature reviewed mostly used probability distribution to represent the relationships among the variables. The use of mathematical expressions and probability notations can be confusing to an audience lacking a mathematical background. Hence, our study adopted the Bayesian network to represent the relationships between the independent and dependent variables in the model to make it easier for audiences to understand.

The identification of three players, namely the adversary, released dataset, and auxiliary datasets that influence disclosure risk, has been supported by Wan et al. (2015), who used game theory to assess the risk of re-identification. The authors adopted game theory to analyze re-identification risk and modeled the data recipient (the adversary) as a player. The adversary was assumed to be intelligent and had access to resources to perform a linkage attack. A linkage attack involves matching key attributes between the released anonymous dataset and auxiliary identified datasets. Wan et al. (2015) implicitly recognize released datasets and auxiliary datasets as players in the disclosure exercise by referring to the linkage attack.

Based on the literature reviewed on identity disclosure risk factors, our study focused on the actual aspects of the players influencing disclosure risk: the adversary, the released dataset, and the auxiliary identified datasets. The study proposes three constructs as predictors of re-identification risk, the analytical competence (AC) of the adversary, the distinguishing power (DP) of the anonymous datasets, and linkage mapping (LM) of the auxiliary identified datasets. The study proposes a conceptual model presented in Figure 1, having not found any studies on the combined effect of the predictors influencing the re-identification risk to fill the gap.

Figure 1

Re-Identification Conceptual Model

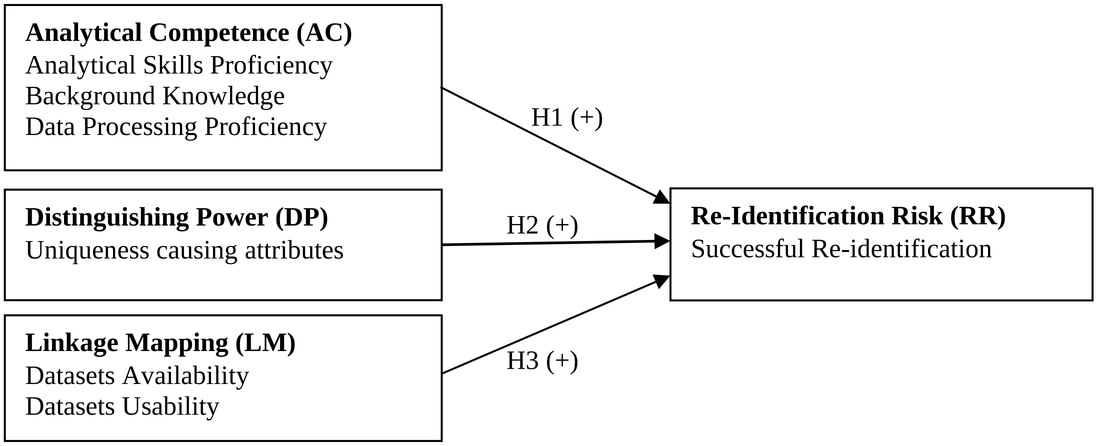

Analytical Competence (AC)
Analytical Skills Proficiency
Background Knowledge
Data Processing Proficiency
H1 (+)
Distinguishing Power (DP)
Uniqueness causing attributes
H2 (+)
Re-Identification Risk (RR)
Successful Re-identification
Linkage Mapping (LM)
Datasets Availability
Datasets Usability
H3 (+)

Figure 1 depicts the three independent constructs hypothesized to influence or predict the dependent variable, the re-identification risk. The independent and dependent constructs, each represented by a rectangle, cannot be measured directly, hence there are manifest or observable variables used to measure them. The manifest variables are listed in each rectangular box.

Hypothesis 1 (H1): The adversary's AC positively influences the re-identification risk.

The main player here is the adversary interacting with the released dataset and seeking to cause disclosure. Most studies focus on background knowledge enabling the adversary to succeed in the re-identification. We argue that the analyst requires more than just background knowledge, and that is how we arrived at the adversary's AC as the factor that influences the likelihood of the adversary succeeding in causing re-identification. We hypothesize that the more AC an adversary possesses, the greater the risk of re-identification.

Reviewed literature pointed to the adversary's background knowledge as a factor influencing disclosure risk (Ganta et al., 2008; Machanavajjhala, 2007; Kifer, 2009; Mohammed et al., 2011; Narayanan & Shmatikov, 2008; Yin et al., 2015). We agree, but argue that background knowledge alone is not sufficient in enabling disclosure. An adversary needs additional skills to interact with datasets, reveal unique records, and then apply background knowledge to cause disclosure. Take the case of the Massachusetts' governor re-identification as an example. The governor was a public figure, and his hospitalization was highly publicized implying an adversary would have background knowledge readily available (Barth-Jones, 2012). But the researcher had to link the released anonymous dataset from the insurance company and that of the voters' register to identify the records - six records had the same date of birth, three were male, and one had a unique zip code (Barth-Jones, 2012). That is how the governor's health record was re-identified (Barth-Jones, 2012; Machanavajjhala et al., 2007; Sweeney, 2002). The researcher must have used data processing and analytical skills to match and establish the governor's record's uniqueness. Mapping the anonymous dataset from the insurance to the identified dataset from the voters' register and getting the matching records required data processing and analytical skills. The initial small set of records identified were then further analyzed using the background knowledge of the adversary (the researcher) to re-identify the governor's record. Establishing record uniqueness from datasets may require dataset operations and manipulations that may not be achievable unless the adversary is proficient in data processing and has analytical skills.

Background knowledge usually goes together with cross-correlation with other datasets to achieve re-identification (Narayanan & Shmatikov, 2008). Hence, data processing and analytical skills are needed to establish cross-correlation among datasets. The disclosure of Netflix users in the famous Netflix Prize competition involved adversaries using data processing and analytical skills to re-identify records (Archie et al., 2018; Narayanan & Shmatikov, 2008; Porter, 2008).

The ability of the adversary to cause disclosure is what we have called AC. This is akin to Xia et al's (2021) modeling the adversary's resources and capabilities. Analytical competence, being a latent construct, cannot be measured directly. We argue that background knowledge lays the ground for or strengthens the adversary's AC . However, background knowledge on its own cannot cause disclosure. The adversary must interact with the datasets, juggle them, make comparisons, establish patterns, draw conclusions, etc. We called these capabilities data processing proficiency. The ability to establish a record's uniqueness, link and match records, and draw conclusions after interacting with datasets requires analytical skills. So, we modeled the adversary's AC as being manifested by analytical skills, background knowledge, and dataset processing proficiency as the measurable variables. Each respondent was categorized based on a 5-point Likert scale for each of the three manifest variables.

Hypothesis 2 (H2): DP of attributes in the anonymized dataset positively influences the re-identification risk.

Distinguishability refers to the characteristics of a dataset that make it possible to identify an individual record uniquely. Depending on the attributes left in the anonymous dataset and the values for those attributes, the adversary may use their analytical skills to make records unique. A unique record combined with background knowledge, or matched with auxiliary identified datasets, could lead to re-identification. Some attributes will have more DP than others, and this may vary from region to region. We hypothesize that the greater the DP of the attributes in the anonymous dataset, the greater the risk of disclosure.

The released anonymized dataset is reported as influencing disclosure risk (Kounine & Bezzi, 2008; Lubarsky, 2017; Winkler, 2005). The anonymized dataset contains both quasi-identifiers and confidential data attributes. Individual quasi-identifiers, such as gender, year of birth, residence, etc., on their own, cannot disclose the identity of any data subject. However, when certain quasi-identifiers are combined, they can facilitate identity disclosure (or re-identification). The combination of certain quasi-identifiers could cause the uniqueness of a record, making such a record distinguishable. Once a record is unique, the adversary can rely on background knowledge or match the record to auxiliary linkage datasets containing explicit identifiers, leading to identity disclosure or re-identification.

In modeling this construct, the focus was on identifying quasi-identifiers that lead to the uniqueness of record(s) and hence are said to have DP . Different geographical regions have different quasi-identifiers that have differing DP s. For example, 63% to 87% of the population in the USA can be re-identified by a combination of gender, year of birth, and zip code (Benitez & Malin, 2010). In other words, the three quasi-identifiers of gender, year of birth, and zip code have high DP in the USA datasets. Another study in the USA reported that 99.98% of USA citizens can be re-identified using fifteen quasi-identifiers (Rocher et al., 2019). However, the quasi-identifiers may vary from one region to another and, more importantly, their DP may vary from one region to another. For example, countries like Kenya without zip codes cannot simply adopt the USA model to establish their disclosure risk. The disclosure risk model has to be domesticated to fit a region (country or state). In particular, a region needs to identify its own quasi-identifiers that have high DP because they enable an adversary to re-identify data. Some regions have regulations on how anonymization should be done before datasets can be released, but other regions do not have established legal frameworks or policies on the anonymization of datasets. With such varying environments, disclosure risk has to be analyzed regionally. In the case reported on in this paper, the respondents provided the attributes they relied on to cause re-identification. Those attributes manifested the DP of the anonymous dataset.

Hypothesis 3 (H3): LM of the auxiliary identified datasets positively influences re-identification risk.

Auxiliary identified datasets do not have confidential or sensitive attributes and are, therefore, considered not to be a threat to the privacy of the data subjects. In some regions, such datasets are shared or made public. Linkage mapping is the ability of such auxiliary datasets to be matched and linked with the anonymized dataset if they happen to have common quasi-attributes. Linkage mapping is, therefore, hypothesized to influence the risk of re-identification positively.

Auxiliary identified datasets contain explicit identifiers that uniquely identify data subjects, such as names, national identity numbers, e-mails, etc., but they have no confidential attributes. In some places like the USA, identified datasets such as voters' registers are readily available and accessible to the public (Barth-Jones, 2012; Benitez & Malin, 2010; Machanavajjhala, 2007; Sweeney, 2002; Xia et al., 2021). Such datasets are released since they do not have confidential information, but they may become auxiliary to the process of re-identification. However, in regions like Kenya, identified datasets are rarely released to the public. This may partly be attributed to the fact that in Kenya, until November 2019, there was no legal framework for data protection and privacy.

The quality of identified datasets enabling the matching of attributes by the adversary is what we named LM. The focus in the literature has only been on the availability or lack of auxiliary identified datasets used for LM purposes (Reiter, 2015; Reiter & Mitra, 2009; Truta et al., 2004). We appreciate the availability of auxiliary identified datasets as a factor influencing disclosure risk, but also argue that the usability of the linkage datasets should not be assumed. The datasets may be available but presented in a format that an adversary may not be able to use. That would hinder the LM, thereby inhibiting disclosure. Therefore, we considered the availability and the usability of identified datasets as the LM measures that influence disclosure risk. The fact that some regions release auxiliary datasets that can cause disclosure while others do not, further support the claim that the re-identification risk has to be regional.

The conceptual model hypothesizes that the positive combined causal effect of the three factors influences disclosure risk and that each factor has predictive strength. Since the re-identification risk is regionally dependent and the Kenyan situation has not been studied, the model was validated using an empirical approach where actual re-identification attempts were made in Kenya.

METHODOLOGY

Our study adopted a quasi-experimental design to study the cause-and-effect relationship between the independent and dependent variables. A quasi-experimental design is used where a true experimental design is not possible for ethical or practical reasons. In a quasi-experimental design, there is no random assignment to treatment. Instead, pre-existing groups are used or a single group serves as both the treatment and control group. This study used a single group serving as both treatment and control group.

Research Sampling

Defining the research dataset is essential for respondents to figure out their working space. For example, the governor of Massachusetts re-identification used the medical records in the health insurance dataset (Dwork, 2011; Rocher et al., 2019; Sweeney, 2002). In the Netflix price completion exercise that ended with disclosure, the dataset was movie ratings (Farzanehfar et al., 2021; Shen, 2013). Several research studies have used clinical datasets (El Emam et al., 2010; El Emam et al., 2011a; El Emam et al., 2011b; Rocher et al., 2019; Taylor et al., 2018). Our study used an educational dataset with two hundred and sixty-six (266) records collected from students at five Kenyan universities. The universities were purposively sampled, but the data subjects, the students, were randomly sampled. Three factors informed the decision to work with university students. The first was that university students are a recognizable constituency, with students coming from all parts of the country. Hence, they are a good representation of the country's general population. Being a recognizable population constituency, the respondents would easily know the cluster of data subjects they would be working with, thus enabling them to narrow their search as they attempt re-identification. The second reason was that university students have good academic knowledge and can be adventurous, making them good at exploring a new field such as re-identification. Using knowledgeable respondents raises the chances of response optimization rather than satisficing (MacKenzie & Podsakoff, 2012). The third reason for choosing university students was legal. In Kenya, most students join the university at the age of eighteen years or older. In Kenya, a person who is eighteen years and above can legally decide to be involved in research. However, a few cases exist of first-year students who are seventeen years old when joining a university. Therefore, only students in their second year of study and above participated in the study as data subjects and respondents. That was to prevent participants who might have been under the legal age of consent from being involved in the study. Therefore, all students who participated in this study were adults, and they willingly consented to participate. The study got authorization from relevant bodies to conduct the research.

The re-identification experiment respondents (the adversaries) were randomly sampled from the five universities and from members of the general public. Some of the respondents had their data in the released anonymous dataset. The respondents' sample size was determined following Cohen's guideline at an effect size of 0.5, a statistical power of 90%, and a confidence level of 95% (Cohen, 1988). Following Cohen's guideline, the minimum sample size was forty-four (44) respondents. However, the study surpassed the minimum sample size and got one hundred and twenty-nine (129) respondents. Thus, the study's high statistical power and high confidence level raised the chances of study findings having high statistical significance and being accurate.

Anonymized Dataset

The collected data were de-identified by removing all explicit (unique) identifiers like names, registration numbers, e-mail addresses, etc. The attributes that were retained in the anonymized datasets were: gender, date of birth, home county, religion, enrolled university, university campus, program enrolled, faculty/school & department, course taken, admission year, current year & semester of study, academic Progress delay, cause of delay, sponsorship type, students sponsorship loan applied, given sponsorship loan, students accommodation type, hobbies. Finally, the anonymous dataset was released to the respondents for them to attempt to cause identity disclosure.

The respondents were given the anonymized dataset and asked to re-identify any data subjects. The respondents then filled in an online questionnaire indicating whether they had managed to re-identify any record or not. Where the respondents claimed to have succeeded in identity disclosure, verification of whether the re-identification was true or false was done. The respondents' skills and attributes relied on during the identity disclosure exercise were obtained through questionnaire feedback. The study used the disclosure attempts' feedback collected from the respondents to evaluate the conceptual model.

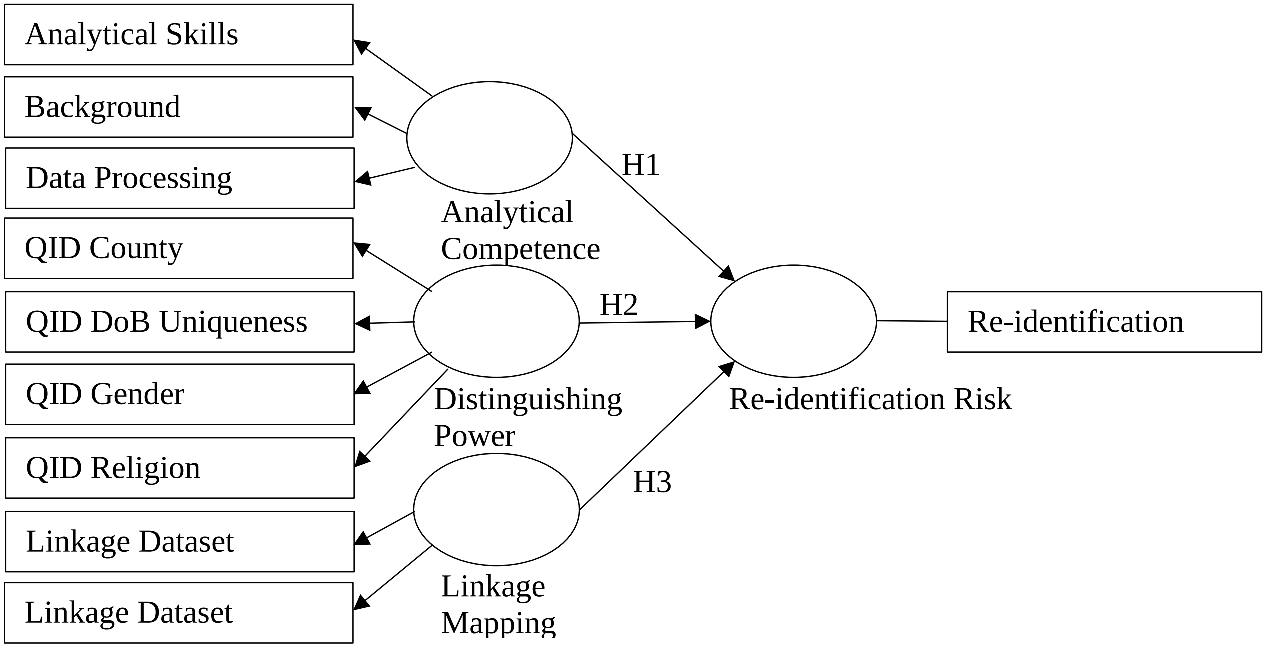
Model Specification and Evaluation

The study used partial least qquares – structural equation modelling (PLS-SEM) to model the re-identification risk. In PLS-SEM, a variable that cannot be measured directly is called a latent variable or construct and is represented using an oval or an ellipse. Each latent construct has measurable or observable variable(s) representing it called manifest variables or indicators, represented using rectangles. The constructs and their respective indicators form the measurement or outer model. The constructs and relationships between them form the structural or inner model. The inner model represents the hypothesized causal relationship. Constructs that do not receive any causal effect from any other but that do have a causal effect influence on another construct are independent or exogenous constructs. A construct that is influenced by another construct is dependent or endogenous. Following the conceptual model in Figure 1, the study used SmartPLS 3.3.7 to evaluate the PLS-SEM model presented in Figure 2. Evaluation of the model is divided into two parts: the measurement or outer model evaluation and the structural or inner model evaluation. The measurement model is evaluated first. If it passes the test, then structural model evaluation is conducted.

Figure 2 shows the four latent constructs; AC, DP, LM, and re-identification risk, represented by ovals or ellipses, and their respective indicator(s), the rectangles, used to measure the construct.

Figure 2

Re-Identification Conceptual Model Specification in PLS-SEM



Note. PLS-SEM = partial least square – structural equation modelling; QID = quasi-identifier; DoB = date of birth.

The AC of an adversary was measured using three indicators: analytical skills proficiency, background knowledge, and data processing proficiency. Each indicator was measured using a 5-point Likert-type scale comprising *Highly Lacking*, *Somehow Lacking*, *Not Sure*, *Somehow in Possession*, and *Highly in Possession*. The DP of the anonymous dataset was measured through four quasi-identifiers (QID): county, date of birth, gender, and religion uniqueness. Respondent's reliance on each QID was measured using a 5-point Likert-type scale comprising *Highly Unrelied On*, *Somehow Unrelied On*, *Not Sure*, *Somehow Relied On*, and *Highly Relied On*. The LM of the auxiliary identified datasets was measured by means of two indicators: datasets availability and datasets usability. Each indicator was measured using a 5-point Likert-type scale comprising *Highly Unavailable/Unusable*, *Somehow Unavailable/Unusable*, *Not Sure*, *Somehow Available/Usable*, and *Highly Available/Usable*. The relationship between the constructs and their indicators is the measurement model. When the arrows are pointing from the construct to the indicators, as in this case, the measurement model is said to be reflective. Otherwise, the measurement model is formative (in this case the arrows point from the indicators toward the construct). When a construct has only one indicator, as in the case of re-identification risk, the relationship is not directional. The relationship among the constructs is called the structural or inner model and is the hypothesized causal relationship being evaluated.

Repeat Experiment

The quasi-experiment in our study underwent a single repeat. In the initial quasi-experiment, the respondents were left to look for auxiliary identified datasets for linkage independently. From that initial quasi-experiment feedback, it was evident that no auxiliary identified datasets were available in the public domain. That necessitated a repeat of the quasi-experiment, in which auxiliary identified datasets for linkage were made available to the respondents.

RESULTS

This section provides the model evaluation results based on the feedback from respondents who took part in re-identification attempts. The evaluation was to check whether the model passed the validity and reliability tests as well as test the hypothesized causal relationship. Therefore, the measurement model evaluation was done first, in which both validity and reliability tests were passed. Finally, the structural model evaluation was undertaken, in which the hypothesized causal relationships were supported. The specifics of the two evaluation results are presented below. The demographics of the respondents who participated in the study are shown in Table 1.

Table 1

Respondents Demographics Information

Demographics		Frequency	Percentage (%)	Cumulative Percentage
Gender	Female	33	25.6	25.6
	Male	96	74.4	100.0
Age in Years	18-25	77	59.7	59.7
	26-35	15	11.6	71.3
	36-45	23	17.8	89.1
	Above 45	14	10.9	100.0
Education Level	Diploma	2	1.6	1.6
	Undergraduate	75	58.1	59.7
	Graduate	2	1.6	61.2
	Masters	43	33.3	94.6
	Ph.D.	7	5.4	100.0

Table 1 summarizes the demographic characteristic of the respondents. There were more male respondents (74.4%) than female respondents. In terms of age, the 18 to 25 age group turned out to be the majority. We attributed this to the involvement in the study of students in universities. The numbers referring to education level corroborate the majority age group. The majority (58.1%) were undertaking their first degree.

Measurement Model Evaluation

The measurement (outer) model evaluation is meant to assess whether the manifest variables (indicators) accurately represent the latent variables (or constructs) they are meant to measure. Being a reflective model, we are required to evaluate the model's indicator reliability, internal consistency reliability, convergent validity, and discriminant validity (Benitez et al., 2020; Hanafiah, 2020). The values to assess each of the four criteria were obtained by running the appropriate PLS algorithms in SmartPLS and are summarized in Table 2 and Table 3.

Table 2 first shows the values of outer loading, being the measure of how well an indicator represents the construct it is meant to represent. For an exploratory and explanatory study as in this case, indicators with outer loading values of 0.4 and above are acceptable (Hanafiah, 2020). Indicators with outer loadings between 0.4 and 0.6 should only be dropped if their dropping improves the constructs' reliability. However, indicators with outer loading values less than 0.4 should be dropped (Hair et al., 2017). On this basis, analytical skills proficiency was dropped as an indicator of AC, leaving background knowledge and data processing proficiency with outer loading values of 0.914 and 0.525, respectively. The outer loading values for manifest variables of LM and DP were above 0.6, making them acceptable. Therefore, the measurement model passed the indicator reliability test.

Table 2

Indicator Reliability and Discriminant Validity

	Indicator Reliability				Discriminant Validity			
	Outer Loadings				Cross Loadings			
Indicators	AC	LM	DP	RR	AC	LM	DP	RR
Background_Knowledge	0.914				**0.914**	0.007	0.334	0.259
Data_Processing_Proficiency	0.525				**0.525**	0.305	0.083	0.123
Linkage_Datasets_Availability		0.842			0.094	**0.842**	0.204	0.192
Linkage_Datasets_Usability		0.905			0.131	**0.905**	0.025	0.243
QID_County_Uniqueness			0.803		0.169	0.009	**0.803**	0.253
QID_DoB_Uniqueness			0.929		0.328	0.128	**0.929**	0.348
QID_Gender_Uniqueness			0.892		0.337	0.170	**0.892**	0.241
QID_Religion_Uniqueness			0.626		0.194	0.080	**0.626**	0.006
Reidentification_Successful				1.000	0.273	0.251	0.326	**1.000**

Note. AC = analytical competence; LM = linkage mapping; DP = distinguishing power; RR = re-identification risk; QID = quasi-identifier; DoB = date of birth.

The other part of Table 2 shows the indicator discriminant validity. Indicator discriminant validity checks whether each indicator represents the construct it is meant to represent better than it would represent any other construct. The indicator discriminant is checked by looking at the cross-loading. The indicator's cross-loading should be higher against the construct it represents than the cross-loading of any other construct. That is shown by the values that are in bold in Table 2 and indeed, they are the highest for the constructs they represented, meaning the indicators passed the discriminant validity test.

The constructs' reliability and validity, yet another measurement model evaluation criteria, are tested using Cronbach alpha, rho_A, composite reliability, and average variance extracted (AVE), as summarized in Table 3.

Table 3 shows the internal consistency or reliability of the constructs using, rho_A, and composite reliability. Hair et al. (2017) state that Cronbach alpha, rho_A, and composite reliability require a threshold of 0.7 and above. However, the composite reliability value is preferred as being more reliable because Cronbach alpha tends to give lower values (Hair et al., 2017). All three independent constructs, i.e., AC, LM and DP, had a composite reliability value of 0.7 and above, hence they all passed the reliability test. According to Hair et al. (2017), an AVE value of 0.5 and above is required for a construct to pass the convergent validity test. AC, LM and DP had values greater than the AVE value of 0.5, passing the convergent validity test. For discriminant validity, all heterotrait-monotrait ratio values were below the threshold of 0.85. Similarly, the Fornell-Larcker criterion indicates the model passed the discriminant validity test. Passing discriminant validity means each construct is distinct from any other.

Table 3

Constructs Reliability and Validity

	Construct Reliability & Validity				Discriminant Validity						
	Reliability			Validity	HTMT			Fornell-Larcker Criterion			
Constructors	CA	rho_A	CR	AVE	AC	LM	DP	AC	LM	DP	RR
Analytical Competence (AC)	0.237	0.314	0.700	0.556				0.745			
Linkage Mapping (LM)	0.695	0.723	0.866	0.764	0.524			0.131	0.874		
Distinguishing Power (DP)	0.857	0.898	0.890	0.674	0.590	0.181		0.321	0.118	0.821	
Re-Identification Risk (RR)					0.521	0.298	0.274	0.273	0.251	0.326	1.000

Notes. HTMT = heterotrait-monotrait; CA = Cronbach alpha; CR = composite reliability; AVE = average variance extracted; AC = analytical competence; LM = linkage mapping; DP = distinguishing power; RR = re-identification risk.

As shown in Table 2 and Table 3, the measurement model passed all the required criteria and could therefore be relied on to assess the structural or inner model.

The data was further assessed to ensure it did not suffer the effects of common measure variance (CMV), which is known to lead to common method bias (CMB) (Min et al., 2016). A CMV is a variance that is associated with a systematic measurement error rather than with the study constructs that the measurement items represent (Min et al., 2016). CMV is further said to be the amount of covariance shared by variables resulting from a common method used during data collection (Malhotra et al., 2006). Studies using self-administered surveys as a data collection method often experience the CMV problem. CMB is reported to have the potential to cause inflation of path coefficients, leading to Type I error, the false positive. It may also cause deflation of the path coefficients, leading to Type II error, the false negative (Kock, 2015).

The study took measures to mitigate CMB effects and confirmed that the data did not suffer from its effects. The first measure that was employed to mitigate the effects of CMV was in sampling, where the respondents were all educated, with the majority (58.1%) being undergraduate students and 33.3% having Masters's degrees, as can be seen in Table 1. That meant the respondents had the capacity to provide accurate responses, which is a quality for respondent optimization when questionnaires are used, as opposed to satisficing (MacKenzie & Podsakoff, 2012).

Secondly, the data was assessed to ensure there were no adverse effects resulting from CMB. One approach used for assessing CMB effects is confirmatory factor analysis, which requires acceptable convergent and discriminant validity tests. Measurement convergent validity is assessed using outer loading, while discriminant validity is assessed using cross-loading (Amora, 2021). The loading and cross-loading obtained are presented in Table 2, and both convergent and discriminant validity criteria were met. Further, convergent validity is established if the constructs have an AVE value of 0.5 and above. As seen in Table 3, all the constructs met this requirement. A second approach used to assess CMB effects was a full collinearity test, which some contend is the preferred method of identifying

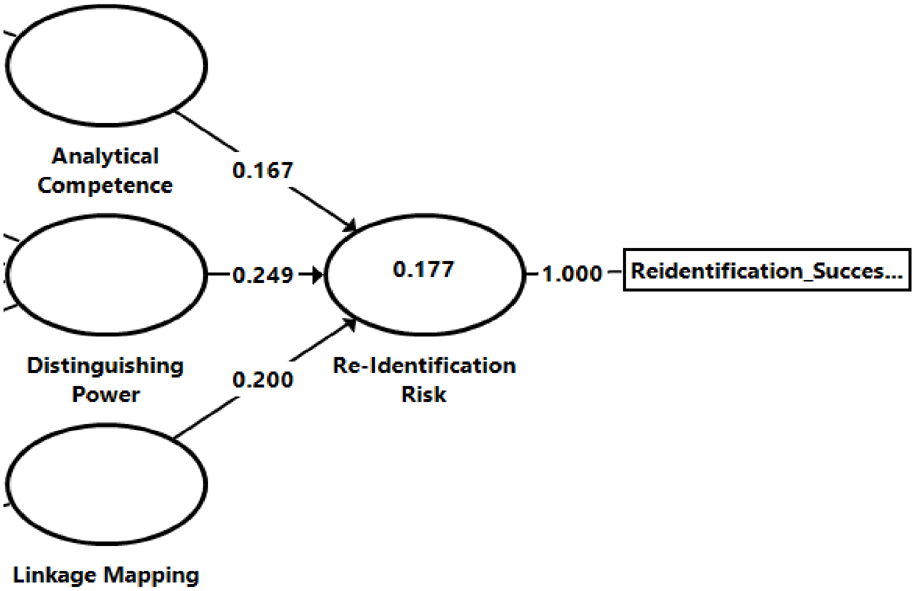
CMB. This procedure gives a variance inflation factor (VIF) for all the constructs in the model. A model is considered free of CMB if it generates VIF values lower than 3.3 for models using classic PLS-SEM. When using factor-based PLS-SEM, the VIF threshold value is 5 (Kock, 2015; Kock & Lynn, 2012). As reported in Table 4, constructs in our study model generated VIF values between 1.024 and 1.126, indicating that the model was free of CMB. Hence, both the confirmatory factor analysis and full collinearity test confirmed that the data did not suffer the effects of common method bias.

Structural (Inner) Model Evaluation

The structural (inner) model evaluation involves examining the path coefficients, testing the model for collinearity, causal relationship significance, coefficient of determination (R^2), effect size (f^2), and predictive power (Q^2). The results are summarized in Table 4. The path coefficient specifies the direction and strength of the causal relationship between the independent and dependent constructs. Figure 3 presents the path coefficients of the model. All the coefficients had positive values, indicating that the predictor constructs positively influenced the re-identification risk. That is in support the hypotheses. The DP of the anonymized dataset had the most predictive strength (0.249), followed by LM of the auxiliary datasets (0.200), and lastly the AC of the adversary (0.167).

Figure 3

Re-Identification Conceptual Model Path Coefficients



Collinearity is a measure of the extent to which the constructs are correlated. It is measured by the VIF. High levels of correlation among constructs mean they do not represent different entities. Hence, a good model should not have a high correlation. VIF values of 5 and above are a sign of high correlation (Hair et al., 2017). The model in our study returned VIF values between 1.024 and 1.126, as shown in Table 4, meaning there were no collinearity concerns for the model. The model passed the collinearity test.

The *t*-values and *p*-values test the significance of the causal relationships between independent and dependent constructs. The causal relationship significance is the test of the hypothesis. A causal relationship is statistically significant if it produces a *t*-value of 1.96 and above and a *p*-value of less than 0.05 (Hair et al., 2017). The study model produced *t*-values of 2.016, 2.814, and 3.399 for AC, LM, and DP respectively (see Table 4). The model also produced *p*-values of 0.044, 0.005, and 0.001 for AC, LM, and DP, respectively, all of which are below 0.05. Therefore, the causal relationships are significant, further confirming the hypothesized causal relationship of the conceptual model.

Table 4

Inner Model Assessment Results

Constructors	Path Coefficient	Collinearity (VIF)	*t*	*p*	f^2	R^2	Q^2
Analytical Competence (AC)	0.167	1.126	2.016	0.044	0.030	0.177	0.134
Linkage Mapping (LM)	0.200	1.024	2.814	0.005	0.048		
Distinguishing Power (DP)	0.249	1.122	3.399	0.001	0.067		

Note. VIF = variance inflation factor.

The coefficient of determination (R^2) is the cumulative measure of the extent to which changes in independent constructs explain changes in the dependent construct. The R^2 is also referred to as the in-sample predictive power and refers to the strength of the predictive model using the data that was sampled. Different study fields have varying threshold levels of R^2, but generally any value above 0.1 is acceptable (Hair et al., 2017). Our study model produced an R^2 value of 0.177, which is slightly above the threshold of 0.1. Hence it is acceptable.

The effect size, denoted as $f^{2,}$ is the measure of change in R^2 when a specific independent construct is removed from the model. If an independent construct has an effect size value less than 0.02, it is said to have no effect (Hair et al., 2017). The constructs in our study model produced f^2 values of 0.030, 0.048, and 0.067, meaning AC, linkage mapping, and DP all had significant effects in causing re-identification.

The last metric for assessing the structural model is the out-of-sample predictive power or relevance, denoted as Q^2. The Q^2 value measures the predictive power of the model using data that is not in the sample. Any value of Q^2 above zero indicates the model is well reconstructed and that the model has predictive relevance (Hair et al., 2017). Our study model produced a Q^2 value of 0.134, meaning the model can be relied on to predict re-identification risk for data other than what was used in the sample. In other words, the model is generalizable.

Since the structural model met all the criteria for assessing it, it is fit for testing the hypotheses. Table 5 summarizes the results of the hypothesis testing. All three causal relationships that the study had hypothesized to influence re-identification risk were supported. Each has a positive influence on re-identification risk but with varying impact strength.

Table 5

Summary of Hypotheses Testing Results

	Hypothesis	Result
H1	Analytical Competence → Re-identification Risk	Significant
H2	Distinguishing Power → Re-identification Risk	Significant
H3	Linkage Mapping → Re-identification Risk	Significant

Limitation

We acknowledge our study's limitations. There could have been some lack of motivation in some of the respondents, leading to some element of satisficing by such respondents. The motivation could have been improved by introducing some form of reward for correctly re-identifying a record. However, the effects of this limitation can be seen to be minimal since CMB was ruled out. Therefore, the limitation did not compromise the quality of the data collected.

DISCUSSION

The identity disclosure risk factors revolve around three players: the released anonymized datasets, auxiliary identified datasets, and the adversary interacting with the datasets. For the released anonymized datasets, the aspect that influences the disclosure risk is the presence of uniqueness-causing quasi-identifiers. Studies have shown that a combination of particular quasi-identifiers can lead to a uniquely identifiable record in the anonymized dataset, leading to identity disclosure due to either the adversary's background knowledge or linkage with auxiliary identified datasets. Identifying quasi-identifiers influencing disclosure risk is essential and is very useful in planning for anonymization. In the USA, for example, a study has shown that 99.98% of Americans can be re-identified using 15 specific demographic attributes (Rocher et al., 2019). This study's model confirmed that the combination of a certain set of quasi-identifiers led to re-identification and identified four attributes (the date or year of birth, gender, religion, and county) that were heavily relied on in causing identity disclosure. The four determined the DP of the anonymized dataset. In this context, the county reflected the geographical residential home place, where Kenya has 47 counties as devolved units. The attributes causing uniqueness should inform data curators doing anonymization to pay close attention to them and to ensure the anonymization minimizes the likelihood of these attributes generating unique records. The suppression approach used for those attributes should ensure that more than one record has the same combined attribute values. Our study acknowledges that the four identified attributes are not the only ones that may jointly identify a single record (cause uniqueness). More research is necessary to identify other sets of attributes capable of causing uniqueness in the Kenyan situation.

Our study revealed the DP of the released anonymous dataset as the factor with the greatest influence on disclosure risk or predictive power. It had the highest path coefficient of all the factors (see Table 4). This is important as data curators should realize that data anonymization needs to be taken seriously as it plays a vital role in determining the disclosure risk.

The person interacting with the datasets (the adversary) is undoubtedly a key disclosure risk factor. Reviewed literature has pointed out background knowledge as the aspect of an adversary that influences the disclosure risk (Ganta et al., 2008; Kifer, 2009; Mohammed et al., 2011; Narayanan & Shmatikov, 2008; Yin et al., 2015; Zhou et al., 2008). The study model corroborated that. But an adversary requires more than background knowledge to cause re-identification. We claim that the AC of the adversary is useful as a re-identification risk predictor. The AC of the adversary will enable them to establish unique records in the anonymized dataset. Only after the unique records have been established can background knowledge become useful in enabling re-identification. However, the adversary must employ data processing skills to establish the unique records in the anonymized dataset. Therefore, the AC predictor is manifested by an adversary's background knowledge and data processing proficiency. This claim was supported by the measurement model, which substantiated that background knowledge and data processing skills were a true reflection of the AC of an adversary. In Table 2, indicator reliability and discriminant validity are evidence of our claim. Then, from reliability validity, convergent validity, and discriminant validity (see Table 3), the hypothesis claiming AC as a predictor of re-identification risk was supported. Further, AC had a *t*-value of 2.016, sufficiently above the threshold of 1.96, and a *p*-value of 0.044, below the threshold of 0.05. The *t*-value and *p*-value showed that the effect of AC as a re-identification risk predictor was statistically significant. The model, however, revealed that the adversary's AC has the least influence or impact of the three disclosure risk factors with a path coefficient of 0.167.

The third factor of re-identification risk relates to the auxiliary identified datasets that were used for linkage purposes with the anonymous datasets. That was named LM . Reviewed literature depicts the availability of the linkage dataset as the indicator of this aspect. We agree with that position. However, we claim that the usability of the identified datasets is equally important. If the identified datasets are presented in a format that the adversary cannot use, they may not be of much help. The study model substantiated the claim with the two indicators having strong reliability and discriminant validity, as shown by the outer loading and cross-loading in Table 2. Similarly, the reliability and validity of the LM construct as a predictor of re-identification produced values of 0.866 composite reliability and AVE of 0.764, both above the 0.7 thresholds. As evidence of significance, the LM construct yielded a 2.814 *t*-value, well above the threshold of 1.96, and a 0.005 *p*-value, well below the threshold of 0.05. That proved that the construct was statistically significant. The LM of the datasets had the second highest path coefficient (Table 4), signifying its impact in influencing the risk of disclosure.

The theoretical approach to estimating the disclosure risk adopts a worst-case scenario, where an adversary is assumed to possess the skills and resources required to enable re-identification. However, this approach tends to overrate the disclosure risk (Xia et al., 2021). The overrated risk then leads to unnecessarily high levels of anonymization, resulting in released datasets with high privacy levels but low analytical utility. Therefore, our study adopted an empirical approach, where data was released to real adversaries in a natural environment. This study's identity disclosure risk estimation model produced a realistic prevailing disclosure risk. That is helpful in avoiding either underestimating or overestimating the risk, both of which affect data privacy and data utility. Having a realistic prevailing risk will lead to anonymization that safeguards both privacy and analytical utility, achieving the balance between the two.

By modeling the combined effects of the three factors influencing a region's disclosure risk, the study revealed that these factors have different influencing impacts. The DP of the released dataset as a result of the retained attributes stood out as having the most significant predictive power, followed by the LM of identified datasets and finally, the adversary's AC .

Theoretical Implications

The theoretical lens for the present study was probability theory, focusing on inter-construct relationships that influence the likelihood of disclosure happening. The current study's findings build on evidence from previous studies that have identified background knowledge, distinguishability in anonymized datasets, and the availability of auxiliary datasets as the influencers of disclosure (Antoniou et al., 2022; Bandara et al., 2020; Domingo-Ferrer & Torra, 2003; Xia et al., 2021). The present study supported the previous results but also brought in new aspects. Firstly, the study introduced data processing proficiency, in addition to background knowledge that was emphasized in previous studies, and formed a new construct called AC of the adversary. The study supported the concept of background knowledge and data processing proficiency as indicators of the adversary's AC, which influences the risk of re-identification. Similarly, the study introduced the usability of the auxiliary datasets in addition to their availability, which was supported in the previous research. The two were used as indicators of a construct called LM, which also influenced re-identification. The research findings have therefore expanded the re-identification risk model.

Secondly, as was stated in previous research, the study supported the view that the distinguishability of attributes of the released datasets influences the re-identification. The new contribution is based on the argument that the risk of re-identification is regionally dependent. That means each region needs to identify the attributes with high distinguishability applicable to the region. The study identified four attributes, namely, date of birth, gender, religion, and county. Those manifested the construct called DP, which influences the risk of re-identification. This finding made the model more specific to the region.

The third theoretical implication of the current study is in providing the predictive strength of each of the three hypothesized relationships. To the best of our knowledge, none of the previous research provided the measure of each predictor as well as a combined effect of the three predictors. It is our view that the quantification of the influence of every predictor is significant, and now we can tell how the influence of each predictor compares to each other.

The fourth implication of the study is in the use of visual Bayesian networks to represent the model. The models encountered in the literature review were mathematical, which may pose interpretation challenges to some audiences. The study has presented a visual model, which is expected to leverage the power of knowledge visualization to effectively communicate to the audience pursuant to knowledge transfer (Van Biljon & Osei-Bryson, 2020).

The findings of the study are in line with one of the streams of Information Communication Technology for Development that study the effects and impacts of Information Communication Technology on developing societies and context (Bon, 2019; Van Biljon & Osei-Bryson, 2020).

Practical Implications

In the absence of an empirical study, the risk of re-identification can only be based on a worst-case scenario that overestimates the risk (Xia et al., 2021). The net effect of overestimated risk is higher levels of anonymization and this in turn affects the utility of the data. Considering that the risk of re-identification is regionally dependent, a realistic assessment of risk of disclosure can only be established using an empirical study. The literature reviewed did not present any such empirical study in the region. The findings of current research are, therefore, likely to impact the practices on data privacy in a number of ways.

Firstly, establishing a realistic disclosure risk is likely to lead to optimal anonymization, which is essential in achieving a balance between preserving data privacy and utility (Asikis & Pournaras, 2020). That would mean that the region could achieve a more useful private data release for data analytics and secondary analysis. An earlier study found the region, Kenya, was not releasing much data (Muturi et al., 2022); this could limit the potential of both data analytics and secondary analysis in the region. The study findings may unlock that potential.

Secondly, the study's findings could inform regional policymakers in formulating an anonymization framework to guide stakeholders in operationalizing Kenya's Data Protection Act 2019 (Republic of Kenya, 2019). Such a framework has the potential to open a new horizon in the area of data release in Kenya, which may spur innovation and knowledge sharing.

CONCLUSION

The study has presented an enriched model for estimating identity disclosure risk by introducing the AC of the adversary, which is manifested by the possession of background knowledge and data processing proficiency. The model also introduced linkage dataset's usability in addition to availability to manifest LM influence on re-identification risk. In addition, we modeled the combined causal effects of the three players in determining the risk. Further, the study has shown how the three factors rank in terms of their predictive power impact, something we had not come across in the reviewed literature. Finally, the model was validated using empirical data in the Kenyan situation, hence domesticating the model.

The enriched identity disclosure estimation model will lead to a realistic estimate of the disclosure risk, informing the right anonymization level. Moreover, the more realistic estimate of the risk will avoid underestimating or overestimating it, making it possible to carry out anonymization to balance privacy and analytical utility.

This study recommends further research on moderating variables' influence on the model. The adversary's motivation to cause disclosure and dataset size have been said to moderate the disclosure risk (Xia et al., 2021). However, as moderating factors, they may only strengthen or weaken the relationships between the constructs. That would not alter the structure of the model. Another area for further research is the need to have a comprehensive list of quasi-identifiers whose combinations lead to disclosure. Further research should explore this area, similar to the USA study that identified the fifteen attributes that can cause 99.98% of Americans to be re-identified (Rocher et al., 2019).

REFERENCES

Amora, J. T. (2021). Convergent validity assessment in PLS-SEM: A loadings-driven approach. *Data Analysis Perspectives Journal*, *2*(1), 1–6.

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. *arXiv preprint arXiv:2203.16921*.

Archie, M., Gershon, S., Katcoff, A., & Zeng, A. (2018). *Who's Watching? De-anonymization of Netflix Reviews using Amazon Reviews*. Technical Report. MIT. Available online: https://courses.csail.mit.edu/6.857/2018/project/Archie-Gershon-Katchoff-Zeng-Netflix.pdf

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

Bandara, P. K., Bandara, H. D., & Fernando, S. (2020, December). Evaluation of re-identification risks in data anonymization techniques based on population uniqueness. In *2020 5th International Conference on Information Technology Research (ICITR)* (pp. 1-5). IEEE https://doi.org/10.1109/ICITR51448.2020.9310884

Barth-Jones, D. (2012). The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now. *SSRN Electronic Journal*, 1–19. DOI:10.2139/ssrn.2076397

Benitez, J., Henseler, J., Castillo, A., & Schuberth, F. (2020). How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research. *Information and Management*, *57*(2), 103168. https://doi.org/10.1016/j.im.2019.05.003

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

Bethlehem, J. G., Keller, W. J., & Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, *85*(409), 38–45. https://doi.org/10.1080/01621459.1990.10475304

Bon, A. (2019, March). Intervention or collaboration? Rethinking information and communication technologies for development. In *6th International Symposium Perspectives on ICT4D: Tackling Global Challenges-Collaboratively*. https://doi.org/10.4324/9780429028236

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 anonymized under EMA Policy 0070 and Health Canada Regulations. *Trials*, *21*(1), 1–9. https://doi.org/10.1186/s13063-020-4120-y

Cohen, J. W. (1988). *Statistical Power Analysis for Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Constitution of Kenya. (2010). Laws of Kenya. *National Council for Law Reporting*, *February*, 191.

Dankar, F. K., El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC Med Inform Decis Making*, *12*(1), 1–15.

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*, 285–293. https://doi.org/10.1016/S0377-0427(03)00643-5

Duncan, G., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, *7*(2), 207–217. https://doi.org/10.1080/07350015.1989.10509729

Dwork, C. (2011). A firm foundation for private data analysis. *Communications of the ACM*, *54*(1), 86–95. https://doi.org/10.1145/1866739.1866758

El Emam, K. (2006). Overview of Factors Affecting the Risk of Re-Identification in Canada. Report, written for Access to Information and Privacy Division of Health Canada, Available online: http://www.ehealthinformation.ca/web/default/files/wp-files/2006-Overview-of-Factors.pdf.

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., Brown, A., Abdelmalik, P., Neisa, A., Walker, M., Bottomley, J., & Roffey, T. (2010). A method for managing re-identification risk from small geographic areas in Canada. *BMC Medical Informatics and Decision Making*, *10*(1),1–13. https://doi.org/10.1186/1472-6947-10-18

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

El 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

El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., & Verma, A. (2011b). The re-identification risk of Canadians from longitudinal demographics. *BMC Medical Informatics and Decision Making*, *11*(1), 1–12 https://doi.org/10.1186/1472-6947-11-46

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

El Emam, K., 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

Elliot, M. (2000). DIS: A new approach to the measurement of statistical disclosure risk. *Risk Management*, *2*(4), 39–48. https://doi.org/10.1057/palgrave.rm.8240067

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

Ganta, S. R., Kasiviswanathan, S. P. S., & Smith, A. (2008). Composition attacks and auxiliary information in data privacy. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '08* (pp. 265–273). Association for Computing Machinery. https://doi.org/10.1145/1401890.1401926

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A Primer on Partial Least Squares Structural Equation Modeling*. SAGE Publications, Inc. https://doi.org/10.1016/j.lrp.2013.01.002

Hanafiah, M. H. (2020). Formative vs. reflective measurement model: Guidelines for structural equation modeling research. *International Journal of Analysis and Applications*, *18*(5), 876–889. https://doi.org/10.28924/2291-8639-18-2020-876

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., & Roth, A. (2014). Differential privacy: An estimation theory-based method for choosing epsilon. In *Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium* (pp. 398–410). IEEE. https://doi.org/10.1109/CSF.2014.35

Jayasinghe, U., Lee, G. M., MacDermott, Á., & Rhee, W. S. (2019). TrustChain: A privacy preserving blockchain with edge computing. *Wireless Communications and Mobile Computing*, *2019*, 1–17. https://doi.org/10.1155/2019/2014697

Kifer, D. (2009). Attacks on privacy and deFinetti's theorem. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (pp. 127-138). Association for Computing Machinery. https://doi.org/10.1145/1559845.1559861

Kniola, L. (2017). Plausible adversaries in re-identification risk assessment. In *Phuse*, *Paper DH09*, 1–10.

https://www.lexjansen.com/phuse/2017/dh/DH09.pdf

Kock, N. (2015). Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of E-Collaboration*, *11*(4), 1–10. https://doi.org/10.4018/ijec.2015100101

Kock, N., & Lynn, G. S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, *13*(7), 546–580. https://doi.org/10.17705/1jais.00302

Kounine, A., & Bezzi, M. (2008). Assessing disclosure risk in anonymized datasets. In *Proceedings of FloCon*, *Section III*.

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

Lubarsky, B. (2017). Re-identification of "anonymized data". Georgetown Law Technology Review. Available online: https://www. georgetownlawtechreview. org/re-identification-of-anonymized-data/GLTR-04-2017

Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M. (2007). *L-diversity: privacy beyond k-anonymity*. *ACM Transactions on Knowledge Discovery from Data*, *1*(1). https://doi.org/10.1145/1217299.1217302

MacKenzie, S. B., & Podsakoff, P. M. (2012). Common method bias in marketing: causes, mechanisms, and procedural remedies. *Journal of Retailing*, *88*(4), 542–555. https://doi.org/10.1016/j.jretai.2012.08.001

McKinsey Analytics. (2016). The age of analytics: competing in a data-driven world. *McKinsey Global Institute Research*. (2016). https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

Malhotra, N. K., Kim, S. S., & Patil, A. (2006). Common method variance in IS research: A comparison of alternative approaches and a reanalysis of past research. *Management Science*, *52*(12), 1865–1883. https://doi.org/10.1287/mnsc.1060.0597

Manrique-Vallier, D., & Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, *107*(500), 1385–1394. https://doi.org/10.1080/01621459.2012.710508

Mello, M. M., Lieou, V., & Goodman, S. N. (2018). Clinical trial participants' views of the risks and benefits of data sharing. *New England Journal of Medicine*, *378*(23), 2202–2211. https://doi.org/10.1056/nejmsa1713258

Min, H., Park, J., & Kim, H. J. (2016). Common method bias in hospitality research: A critical review of literature and an empirical study. *International Journal of Hospitality Management*, *56*, 126–135. https://doi.org/10.1016/j.ijhm.2016.04.010

Mohammed, N., Chen, R., Fung, B. C. M., & Yu, P. S. (2011). Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11* (p. 493–501). Association for Computing Machinery. https://doi.org/10.1145/2020408.2020487

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

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings - 2008 IEEE Symposium on Security and Privacy* (pp. 111–125). IEEE. https://doi.org/10.1109/SP.2008.33

Nelson, G. S. (2015). Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification. In *SAS® Global Forum proceedings* (pp. 1-23). SAS Institute Inc. https://support.sas.com/resources/papers/proceedings15/1884-2015.pdf

Orooji, M., & Knapp, G. M. (2018). Improving suppression to reduce disclosure risk and enhance data utility. In K. Barker,

D. Berry, & C. Rainwater (Eds.), *IISE Annual Conference and Expo 2018: Proceedings* (pp. 13-18). Institute of Industrial & Systems Engineers (IISE).

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics*, *6*(4), 487–500. https://doi.org/10.1080/07350015.1988.10509697

Patel, K., & Jethava, G. B. (2018, April). Privacy preserving techniques for big data: A survey. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 194-199). IEEE. https://doi.org/10.1109/ICICCT.2018.8473289

Personal Data Protection Commission Singapore (2018). *Guide to Basic Data Anonymisation Techniques*. Available online: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf

Porter, C. C. (2008). De-identified data and third party data mining: The risk of re-identification of personal information. *Shidler JL Com. & Tech.*, *5,* 1.

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*, *50*, 1299–1323. https://doi.org/10.1007/s11747-022-00845-y

Reiter, J. P. (2015). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, *100*(472), 1103–1112.

Reiter, J. P., & Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, *1*(1), 99–110. https://doi.org/10.1198/016214505000000619

Republic of Kenya. (2019). The Data Protection Act. The Office of the Data Protection Commissioner (ODPC). https://www.odpc.go.ke/dpa-act/

Ribeiro, S. L., & Nakamura, E. T. (2019). Privacy protection with pseudonymization and anonymization in a health IoT system: Results from OCARIoT. In *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019* (pp. 904–908). IEEE. 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*, 3069. https://doi.org/10.1038/s41467-019-10933-3

Santu, S. K. K., Bindschadler, V., Zhai, C., & Gunter, C. A. (2018, January). NRF : A naive re-identification framework. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society (WPES'18)* (pp. 121–132). Association for Computing Machinery. https://doi.org/10.1145/3267323.3268948

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. Available online: https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF

Shen, E. (2013). Privacy-preserving and usable data publishing and analysis [Doctoral dissertation, North Carolina State University]. North Carolina State University Research Repository. https://repository.lib.ncsu.edu/handle/1840.16/8605

Shlomo, N. (2010). Releasing microdata: Disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality*, 2(1). https://doi.org/10.29012/jpc.v2i1.584

Shlomo, N., & Skinner, C. (2022). Measuring risk of re-identification in microdata: State-of-the art and new directions. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *185*(4), 1644–1662. https://doi.org/10.1111/rssa.12902

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 minimizing re-identification risk in research data derived from health care records. *EGEMs*, *7*(1), Article 6. https://doi.org/10.5334/egems.270

Simon, G., Shortreed, S. M., Yates Coley, R., Iturralde, E. M., Platt, R., Toh, S., & Ahmedani, B. (2020). *Toolkit for Assessing and Mitigating Risk of Re-identification when Sharing Data Derived from Health Records*. Available online: https://www.sentinelinitiative.org/sites/default/files/Methods/Sentinel_Report_Toolkit-Assessing-Mitigating-Risk-Re-Identification-Sharing-Data-Derived-from-Health-Records.pdf

Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society*, *64*(4), 855–867.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, *671*(2000), 1–34.

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. *Technology Science*, *2017:20170*.

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Taylor, L., Zhou, X. H., & Rise, P. (2018). A tutorial in assessing disclosure risk in microdata. *Statistics in Medicine*, *37*(25), 3693–3706. https://doi.org/10.1002/sim.7667

Truta, T. M., Fotouhi, F., & Barth-Jones, D. (2004). Disclosure risk measures for the sampling disclosure control method. In *Proceedings of the 2004 ACM Symposium on Applied Computing* (pp. 301–306). Association for Computing Machinery. https://doi.org/10.1145/967900.967964

van Biljon, J., & Osei-Bryson, K. M. (2020). The communicative power of knowledge visualizations in mobilizing information and communication technology research. *Information Technology for Development*, *26*(4), 637–652. https://doi.org/10.1080/02681102.2020.1821954

Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E. W., Kantarcioglu, M., Ganta, R., Heatherly, R., & Malin, B. A. (2015). A game theoretic framework for analyzing re-identification risk. *PLoS ONE*, *10*(3), e0120592. https://doi.org/10.1371/journal.pone.0120592

Winkler, W. E. (2005). Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Statistics*, *9*, 1-14.

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.

Yin, L., Wang, Q., Shaw, S. L., Fang, Z., Hu, J., Tao, Y., & Wang, W. (2015). Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PLoS ONE*, *10*(10), e0140589. https://doi.org/10.1371/journal.pone.0140589

Zhou, B., Pei, J., & Luk, W. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations Newsletter*, *10*(2), 12–22. https://doi.org/10.1145/1540276.1540279

Zlatolas, L. N., Hrgarek, L., Welzer, T., & Hölbl, M. (2022). Models of privacy and disclosure on social networking sites: A systematic literature review. *Mathematics*, *10*(1), 146. https://doi.org/10.3390/math10010146