Anonymization of data for open science in psychology: Part I — traditional anonymization

Jiří Novák Matthias Templ Carolin Strobl

October 14, 2024

Abstract

Psychology as a field experienced a crisis caused by a lack of reproducibility. On the one hand, this gave rise to distrust in the results, but on the other hand, it enabled the development of open science practices. A key component of open science is the dissemination of well-anonymized open data, which facilitates transparency while protecting privacy.

More openly available data would make research more transparent and accessible. Unfortunately, many datasets cannot be available even to other researchers for privacy reasons. Despite these challenges, researchers are increasingly expected to make data available for review, reanalysis, and reuse.

In this paper, we present good practices of statistical disclosure control for psychologists. The practices are divided into two separated parts: the first part consists of traditional approaches, and the second part focuses on the modern approach of using synthetic data. The traditional approaches modify data so that it can be disseminated without revealing confidential information that may be associated with specific respondents.

The paper seeks to provide practical insights into how statistical disclosure methods can effectively balance the need for data transparency and privacy in psychological research. Through a detailed case study, we demonstrate the practical application of these methods to protect sensitive data.

Keywords: open science, confidentiality, reproducibility, anonymization, sdc

1 Introduction

Open science practices are gaining importance in contemporary research, particularly in psychology, where researchers handle susceptible data. The sensitive nature of psychological variables often hinders data sharing, contributing to the replicability crisis in the field.

Therefore, data cannot be shared well among researchers, which is one reason for the crisis of scientific replicability. Data from publicly funded research should be more widely disseminated, at least among scientists. They should then be able to replicate studies, try new methods on the given data, or verify the results presented in the given article or study.

Open science is a movement that has been gaining strength and importance in recent years. The movement aims to make scientific research funded by public resources openly available so that it can be reused, replicated, traced, and trusted. This transparency also enhances the financial efficiency of research, fostering better global scientific collaboration.

All started by Budapest Open Access Initiative [12] in 2002, which was then supplemented with a set of rules in 2012 [12] and 2022 [13]. This was followed by the Bethesda Statement on Open-Access Publishing [22] in 2003 and Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities [21].

Budapest Open Access Initiative (BOAI) [11] define *Open Access* (OA) as *"free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself".*

accurate data.

Lastly disclosing information that can be linked to specific individuals or entities is unethical. Declaration on Professional Ethics [10] set of Ethical Principles for statisticians and a wide array of creators and users of statistical data and tools. Disclose information that can be directly or indirectly linked to specific individuals or entities without their consent is considered unethical. Such actions may compromise privacy, lead to potential misuse of data, and violate principles of confidentiality. Ethical considerations require careful handling of sensitive information to prevent harm and uphold respect for personal and organizational boundaries. Necessary steps must be implemented to ensure that data are released in a way that protects the confidentiality of individuals, preventing their identities from being disclosed or inferred.

The replication and reproducibility crisis has received much attention in the last decade [28]. The replication crisis is a phenomenon in science, particularly in psychology and medicine, where many previously published scientific studies cannot be replicated or reproduced with similar results. This means that when other researchers attempt to repeat the experiments using the original methodology, they fail to achieve the same outcomes. This issue raises concerns about the reliability and validity of scientific findings, leading to criticism of certain research practices, such as statistical errors, flawed experimental design, or the publication of only positive results.

In 2015, Open Science Collaboration [18] of 271 authors examined the reproducibility of experiments in psychology. They selected about 100 studies from three psychology journals with the aim of achieving the same results as the original studies. Only 36% of original studies achieved significant results.

X

The aim of this paper is to explore and elucidate the various statistical disclosure control methods available for the anonymization of data in the context of open science within psychology. This paper will specifically focus on presenting a detailed examination of traditional anonymization, demonstrating their application to a psychological dataset. Through this analysis, the paper seeks to provide practical insights into how these methods can be effectively utilized to balance the need for data transparency and privacy in psychological research.

2 Disclosure risk

Disclosure is defined [9] as when a person or an organisation (intruders) recognises or learns something that they did not already know about another person or organisation via released data. The first step of anonymization is evaluating risks that threaten the data. This is approached by creating disclosure scenarios [9] tailored to the data that will be disseminated. The scenarios depend on the intruder's intentions, prior knowledge, and the available data. Depending on the intention of the intruders, their type of a priori knowledge and the microdata available. Disclosure risks that may occur are identity disclosure, attribute disclosure, defined by [7], and inferential disclosure, defined by [5]. Identity disclosure is the association or linking of the specific respondent to a record. Preventing this is of great importance. Attribute disclosure occurs when an intruder successfully links a record to a respondent. Inferential disclosure happens when a data user deduces new information about a respondent from the published data.

Dissemination of findings from psychological science is discussed by Purtle [20], they advocate for a structured, evidence-based approach to effectively communicating psychological research by conducting audience research, segmenting target groups, and testing dissemination strategies, with attention to personalization and privacy concerns. However, the article does not provide detailed guidance on how to ensure privacy during the dissemination process.

Couper [4] examined how perceptions of privacy, confidentiality, or risk of data disclosure affect individuals' willingness to participate in surveys. The findings show that while objective disclosure risk does not significantly reduce survey participation, perceptions of risk and topic sensitivity substantially lower the willingness to participate. Respondents are more likely to decline participation in surveys on sensitive topics, driven by concerns about privacy and potential harm, rather than the actual probability of data disclosure.

** Suggestion:

2.1 : General discussion on disclosure risk.

Say more on
- Identity disclosure + example
- Attribute disclosure + example
- Inferential disclosure + example
- Membership disclose + example

I described one in the paper

~~2.2 ...~~

2.2. Attacker scenarios
- Nosy neighbor + example + risk
- Linkage attack + example + risk
- Outlier attack
- Differential Attack
- Homogeneity Attack
- Inference attack

I wrote one in the paper

⇒ table ?

2.3. Kind of data

*** - survey data and questionnaire data
  - type
  - role
  - data structure
  - anonymisation needs
- Qualitative data
- Psychometric Test Score
- Longitudinal data
- Biometric and physiological data
- Behavioral data (eye tracking, ...)
- Genetic data
- Social network data

I wrote one example in the paper

*** Data specialities regarding SDC

- highly granular, sensitive item scale
- small sample size and unique groups
- multivariate correlations and dependence
  (psych. scales are often highly correlated...)
- presence of sensitive variables
- response patterns and profiling?
- longitudinal data and tracking
- high utility requirement for psychological constructs

Text + Table ?

24. Evaluating disclosure risk

- Uniqueness analysis : + Example
- k-anonymity
- l-diversity, t-closeness, ..
- risk scores
- ~~idk~~
- individual risk when survey samples

Last but not least, Kilovaty [14] describes that the disclosure of sensitive data can lead to profound emotional and psychological consequences, including heightened anxiety, depression, post-traumatic stress disorder, and a range of other mental health challenges that can severely impact the well-being of those affected.

Wairimu [26] collected real-world examples of privacy breaches in healthcare. These include a case where a nurse in Florida disclosed a woman's medical records, causing fear and embarrassment. The nurse's actions, which involved sharing sensitive information with unauthorized individuals, highlighted the dangers of improper access to personal medical data[1]. In the case of Hinchy v.Walgreen Co., a patient's prescription history was shared with her ex-boyfriend, leading to emotional distress. The court found Walgreens liable for both HIPAA violations and negligence, resulting in a $ 1.44 million judgment against the company[2]. Hackers leaked the medical data of high-profile athletes from the World Anti-Doping Agency, exposing sensitive medical information about the athletes and potentially causing distress for the athletes involved[3]. One patient's HIV status was publicly accessible for months, which occurred when her medical information was shared inappropriately. This breach not only caused emotional harm but also led to feelings of fear, stigma, and a profound loss of trust in the healthcare system, leading to significant emotional harm[4]. One victim of medical identity theft incurred nearly $20,000 in fraudulent bills, causing financial strain and distress. It was caused by hackers stealing medical records and selling them on the dark web. These data often include sensitive personal information like Social Security numbers, birth dates, and medical histories[5]. An NHS staff member unlawfully accessed and shared a relative's confidential medical records with other family members, which resulted in psychological harm and medication use. The case emphasizes the serious consequences of mishandling personal medical data, even within families[6]. In Finland, psychotherapy patients were subjected to blackmail after a data breach. The hackers obtained highly confidential patient records, including details of therapy sessions, and then attempted to blackmail both the patients and the psychotherapy clinic. Victims received ransom demands threatening to publicly release their personal mental health information if payments were not made. The breach caused widespread distress among patients and raised concerns about the security of sensitive medical data in the healthcare sector.[7]. In a UK case, women were stalked after the unauthorized access occurred following a data breach at University Hospital Crosshouse in Scotland, where the woman's details were improperly shared[8].

Walsh [27] reviewed privacy risks associated with sharing clinical data in psychological and psychiatric research, particularly identity, attribute, and membership disclosure. The authors recognize that identity disclosure is a significant risk when sharing clinical data. Such disclosures can occur, for example, when the remaining information in a patient's record is connected to another source that reveals their identity. These sources might be publicly available or accessible only to a limited group, such as neighbours, friends or family members with whom the patient has shared their involvement in a research study. So even when explicit identifiers (like names or Social Security numbers) are removed, individuals can still be re-identified by combining other seemingly innocuous details (such as demographics, location, or medical history) with publicly available information, like voter registration records. They highlight that this risk is particularly concerning in the field of psychology and psychiatry because of sensitive data, like psychiatric diagnoses or treatments. As mentioned in Wairimu [26] and Kilovaty [14], disclosed diagnoses or treatments could lead to embarrassment or result in stigmatization or discrimination by family, friends, the wider public, or organizations that might misuse this information in ways that could harm the individual's well-being. The authors emphasize that attribute disclosure can happen even when identity disclosure has been prevented, making it a complex issue to address when sharing clinical data. To illustrate, consider that a patient's identifiable details (such as demographics) are combined to resemble those of many others in a study on paranoid schizophrenia. If everyone in the study shares the same diagnosis and someone can identify that a specific individual is part of the study, they will be able to deduce that this person has been diagnosed with the disorder. This

[1] https://www.tampabay.com/archive/2013/06/29/records-breach-lets-out-secret/

[2] https://www.bswllp.com/the-intersection-of-hipaa-and-negligence-pharmacists-violation-cost-walgreens-144-million

[3] https://www.bbc.com/news/world-37369705

[4] https://www.npr.org/sections/health-shots/2015/12/10/459091273/small-violations-of-medical-privacy-can-hurt-patients-and-corrode-trust

[5] https://www.cbsnews.com/news/hackers-steal-medical-records-sell-them-on-dark-web/

[6] https://www.hayesconnor.co.uk/news-resources/case-study/nhs-family-member-shared-confidential-medical-information/

[7] https://www.theguardian.com/world/2020/oct/26/tens-of-thousands-psychotherapy-records-hacked-in-finland

[8] https://www.cumnockchronicle.com/news/17310994.stalker-rap-hospital-data-breach/

is also called a membership disclosure. Membership disclosure is another form of privacy violation, highlighted when researchers showed that genomic summary statistics from case-control studies could allow someone with access to a person's DNA data to confirm their participation in a study. According to [23], membership disclosure is a particular case of attribute disclosure.

3 Overview of SDC Methods

Statistical Disclosure Control (SDC), as defined by [9], seeks to protect statistical data in a way they can be released without revealing confidential information that can be linked to specific individuals. The goal of SDC methods is to find an optimal solution for both the risk of disclosure and the utility of protected published data.

Methods intended to protect microdata are described in detail in the publications of Hundepool [9]. In general, SDC methods can be divided according to when they are applied. The method can be applied directly to microdata, then we talk about pre-tabular methods, or to aggregate data in tables or hypercubes, and then we talk about post-tabular methods. The methods applied to microdata are naturally all pre-tabular methods. We further distinguish the methods of modifying the values into three main groups: non-perturbative methods, perturbation methods, and methods of creating synthetic and hybrid data. Non-perturbative methods adjust the detail of the data display, perturbative methods add noise to the data, and synthetic and hybrid data generation methods generate new data based on the original data.

3.1 SDC stages

For the purposes of psychologists, we simplify the process of SDC described in [9]. This process consists of a few essential stages.

Stage 1: Assess the need for confidentiality protection Stage 2: Key characteristics and use of data Stage 3: Disclosure risk Stage 4: Disclosure control methods Stage 5: Implementation

3.2 Disclosure scenario

From these scenarios, the types of data disclosures emerge.

At the beginning of the SDC process, data must be cleaned of direct identifiers such as name, social security number or similar ID, address, and e-mail. This is called de-identification [6] or pseudo-anonymization. This alone does not prevent de-identification and revealing new information to the intruder.

X

X

X

X

4 Case Study: Anonymizing a data set from psychology

x

4.1 Data

The data for this example is from the Answers to the Machivallianism Test, a version of the MACH-IV from Christie and Geis [2], which comprises 73,489 records. The dataset includes both Likert-rated items and demographic variables.