PREPRINT (to appear in PETS '27): An Exploratory User Study of Disclosure Avoidance Tools for Scientific Microdata

Wentao Guo
University of Maryland
wguo5@umd.edu

Wellington Esposito Barbosa
The George Washington University
wellingtone@gwu.edu

Adam J. Aviv
The George Washington University
aaviv@gwu.edu

Michelle L. Mazurek
University of Maryland
mmazurek@umd.edu

Abstract

Social, behavioral, and health scientists are often required or encouraged to de-identify and publish data to support open science. Prior work shows that scientists, lacking support, avoid systematic disclosure avoidance methods that could help them achieve their own privacy goals more rigorously. Better tools could help but must be tailored to scientists' real-world needs in order to have an impact. To understand these design needs, we conducted an exploratory user study in which experienced researchers used one of two disclosure avoidance tools on a real research dataset. We highlight two main findings. First, a high degree of skepticism about disclosure risk in the original data led participants to perceive these tools as extreme. Second, these tools contain promising design patterns for helping scientists understand impacts on data utility, but they largely fail to improve scientists' understanding of privacy. Our findings indicate a need for better approaches to evaluate disclosure risk, and we provide a foundation for tailoring disclosure avoidance tools to the needs of researchers participating in open science.

Keywords

disclosure avoidance, de-identification, data privacy, differential privacy, synthetic data, *k*-anonymity, open science, user study

1 Introduction

Social, behavioral, and health scientists are often required or encouraged to publish data to support open science [3, 25, 46, 50, 75, 79, 80]; as a result, tens of thousands of datasets about human subjects are available in online repositories [24, 29]. To prevent the disclosure of private information about research participants, scientists de-identify data by removing or modifying names, demographics, and other potential identifiers. Unfortunately, disclosure avoidance[1] is a difficult task for which most scientists lack adequate support.

First off, there is no singular ideal solution. Privacy researchers have found vulnerabilities in heuristic de-identification methods again and again [10, 51]; meanwhile, differential privacy [13] provides provable protections, but many deployments (in order to retain data utility) have ended up with arguably trivial privacy guarantees[2] while simultaneously struggling to gain acceptance among scientists [42, 61]. Further, fundamentally difficult choices are exacerbated by insufficient support for scientists in practice. Our own prior work found that available de-identification guides often make misleading claims while also neglecting to provide good instructive examples [20]. Through interviews, we found that scientists have almost no exposure to systematic disclosure avoidance frameworks such as *k*-anonymity and differential privacy, and their methods are often at odds with their de-identification goals [21].

Well-designed disclosure avoidance tools could help scientists on many fronts: unlocking more robust privacy techniques, providing better feedback on impacts to utility, and reducing time and error. To achieve these benefits in practice, though, tools must provide information and workflows that enable users who are not experts in disclosure risk to de-identify data and reason about whether the output is acceptable for secondary use. As such, disclosure avoidance tools for scientists likely have different design requirements than tools for privacy experts. This study explores the design of disclosure avoidance tools for scientists. We focus particularly on microdata—datasets with line-by-line information about individuals, rather than aggregate statistics—because publishing microdata is overwhelmingly the norm in open science. While valuable work has focused on designing tools to protect statistics using differential privacy [40, 62, 66], we believe that meeting scientists where they currently stand is complementary and necessary.

The design of disclosure avoidance tools for microdata is largely unexplored in human-centered computing research. As a first step, we conducted a user study with two real-world tools that could help scientists de-identify data more systematically. We aim to learn about tool design at a high level, focusing on users' interactions with different *kinds* of features and workflows, rather than low-level concerns such as the placement of interface elements. This framing enables us to generate findings that transfer better to future tools, which will likely differ radically from current ones. Nevertheless, we believe that interacting with real tools is crucial—despite notable usability and privacy issues (see §2.3)—because it forces participants to grapple directly with the kinds of trade-offs that necessarily result from systematic disclosure avoidance methods.

[1] In this paper, we use the terms *disclosure avoidance* and *de-identification* interchangeably to describe a broad range of approaches that "[remove] the association between a set of identifying data and the data subject," as defined by NIST [18].

[2] For example, the 2020 U.S. Census was released with ϵ of 19.61 [76], while two synthetic datasets on human trafficking [74] and live births [22], respectively, used ϵ of 12 and 9.98. These are large ϵ; as Dwork et al. write, large ϵ are "difficult to reason about" and may offer "little to no protection" [14].

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit https://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.
Proceedings on Privacy Enhancing Technologies YYYY(X), 1–21
© YYYY Copyright held by the owner/author(s).
https://doi.org/XXXXXXX.XXXXXXX

We recruited 23 social, behavioral, and health scientists for our study. Each participant de-identified a real research dataset using one of two disclosure avoidance tools: sdcMicro [73] (k-anonymity), or SmartNoise [65] combined with SDMetrics [11] (differentially private synthetic data). We investigate three research questions:

RQ1: How do scientists perceive the acceptability of systematic disclosure avoidance tools, after trying them? Existing tools, including the two in this study, do little to help users understand baseline disclosure risk; accordingly, many participants believed the original data was already well protected, leading them to conclude that the tools are extreme. In general, participants were cautious of overreliance on technical methods to protect data, raising concerns around data utility, transparency, and scientific norms. At the same time, for some participants, firsthand usage allayed initial doubts and led to cautious optimism about tools' usefulness.

RQ2: How do scientists interpret tools' metrics and visualizations for evaluating data utility and privacy? We find that tools combine utility metrics and visualizations in useful ways, forming positive feedback loops: metrics helped participants decide where to focus their attention when interacting with visualizations, and visualizations in turn helped build participants' understanding of metrics—though some important nuances were obscured. However, participants struggled to make sense of privacy information, indicating a need for improved visualizations and examples to help users build up their intuition about privacy metrics.

RQ3: What workflows should disclosure avoidance tools support to meet scientists' needs? We report participants' perspectives on two main tool workflows that they would like to see developed further: providing better support for decision-making, and generating documentation that can inform secondary users about the uses and limitations of the de-identified data. We also describe differing perspectives on whether tools should give scientists greater control over how disclosure avoidance algorithms are applied.

2 Background

Here we summarize the state of disclosure avoidance in open science. We also describe existing tools and some of their limitations.

2.1 Disclosure avoidance in open science

Social, behavioral, and health scientists are increasingly required or expected to publish research data by the institutions that fund, support, and publish their work. As discussed in prior work, this has resulted in tens of thousands of human-subjects datasets published in repositories online, often containing sensitive information about individuals [21]. Though repositories attempt to guard against misuse of data through access and use restrictions, these methods are fallible. In a recent example, fringe race scientists allegedly lied about their research intentions in order to obtain genetic and health data from the National Institutes of Health [33].

To mitigate privacy harms, scientists are often expected to blur or remove details in their dataset that could expose individuals' identities or sensitive information, while also retaining as much of the data's utility for secondary use cases as possible. We refer to this process interchangeably as *disclosure avoidance* and *de-identification*. Carvalho et al. [8] provide a detailed overview of privacy risks and disclosure avoidance methods for *microdata*—datasets that contain discrete records for individuals, rather than aggregate statistics.

There are serious challenges to de-identifying scientific data in practice. In prior work, we analyzed 38 de-identification guides, finding that they convey limited threat models and disclosure avoidance methods, and that they provide few examples of how to de-identify data systematically [20]. We also interviewed 24 scientists with de-identification experience, finding that they typically employ informal disclosure avoidance methods that are likely to result in higher disclosure risks than aimed for [21].

Systematic disclosure avoidance methods like k-anonymity [69] and differential privacy [13] could improve protections over current norms. For k-anonymity, one specifies a set of identifying variables that an attacker could leverage; then, data is coarsened or deleted to ensure each record belongs to an anonymity set of at least k records. For example, in a dataset that is 5-anonymous with respect to gender, age, and state, if a record describes a man from Maryland who is 19, then there must be at least four other records of 19-year-old men from Maryland. In a recent case study on de-identifying medical data, Carvalho et al. used k-anonymity as an intuitive way to measure reduction of disclosure risk [7]; at the same time, k-anonymity has notable limitations as a privacy metric (see §2.3).

To achieve differential privacy on microdata, one typically uses a randomized algorithm to generate synthetic data with similar properties to the original. Differential privacy formally limits how much the synthetic data tends to change when any one individual's data is altered, guaranteeing a limit to the privacy risk of being in a dataset. The level of privacy is controlled by the parameter ϵ: smaller values provide a stronger privacy guarantee but tend to produce less similar data, while larger values provide weaker privacy and greater similarity. Several recent works have tested differentially private synthetic data generation methods on real human-subjects datasets, often demonstrating that they can support specified use cases—but also that the utility trade-offs can be significant and can vary in complex ways depending on dataset characteristics, synthetic data algorithm, and parameter choices [17, 31, 55, 59, 67, 78].

2.2 Disclosure avoidance tools

Tools have been developed to assist with disclosure avoidance for structured microdata. sdcMicro [73] and ARX [57] are notable tools with graphical user interfaces that enable users to delete, coarsen, or perturb data under frameworks such as k-anonymity and differential privacy. Some tools generate synthetic data: SmartNoise [65] and the Synthetic Data Vault [12] are examples of tools that can synthesize data with differential privacy guarantees. In Section 7.2 of their survey paper, Carvalho et al. describe existing disclosure avoidance tools for microdata in greater detail [8].

Tools can also help users evaluate the utility and privacy of de-identified data. For example, sdcMicro and ARX provide counts of deleted values, metrics of statistical similarity, and visualizations of changes in marginal distributions. Similarly, SDMetrics [11] and synthpop [49] provide metrics and visualizations to compare any synthetic dataset to its original. In contrast, SDNist is a project to

benchmark synthetic data algorithms based on pre-determined, dataset-specific utility considerations and disclosure risk models [71]. As such, SDNist currently works with just two U.S. Census datasets, but it is an interesting example of how users could be guided through utility and privacy evaluations.

Tools can assist with disclosure avoidance in other contexts, such as redacting free text [35, 77] and releasing statistics [54]. Sarathy et al. conducted a study similar to ours, in which scientists and repository curators used a tool to produce differentially private statistics; participants had difficulty understanding the implications of choices such as parameter selection, and they were concerned about losing the ability to conduct exploratory data analyses and replication studies [62]. In parallel, Nanayakkara et al. developed a tool to help scientists understand utility and privacy trade-offs of ϵ values for differentially private statistics [40]. While important, these other contexts are out of scope for our work, which focuses on disclosure avoidance for microdata because it is the norm in open science—and likely to remain so for the foreseeable future.

Recently, some have explored using AI to help users reason about disclosure risks and mitigation strategies. sdcMicro recently introduced AI features to suggest classifications for variables (e.g., as identifiers or not) and disclosure avoidance steps [72]. Zheng et al. developed a technique using LLMs to estimate the total number of people who match a given set of attributes (i.e., the global k value); in empirical tests, their technique performed better than other LLM-based approaches, though predictions still differed around 30% of the time by a factor of 5 or more from estimates based on real statistics [82]. As of now, AI-assisted disclosure avoidance tools have barely been explored and are essentially untested, but initial developments suggest interesting directions for future work.

2.3 Limitations of disclosure avoidance tools

In practice, while some scientists use home-brewed tools to scaffold disclosure avoidance, it is nearly unheard of to use tools to carry out data transformations [21]. This is perhaps just as well, since existing tools are not necessarily mature or well-tailored to scientists' needs.

Usability (especially for non-experts) is a pervasive issue. In a user study of differential privacy tools, participants made implementation mistakes due to issues such as unclear documentation and inadequate support for complex tasks like calculating sensitivity [47]. There is a lack of good usability studies on disclosure avoidance tools for microdata, but we attest that we encountered many issues when assessing candidate tools for this study. To provide just a few examples, ARX and Amnesia have unintuitive interfaces for creating generalization hierarchies; labels on sdcMicro visualizations overlap; SDMetrics visualizations fail to accommodate red-green color blindness; and ARX often refers users to academic papers in order to understand its utility and privacy metrics.

Tools also have privacy limitations. Several prominent tools use k-anonymity, which does not prevent disclosure of sensitive values. More complex frameworks extend k-anonymity in an attempt to address this issue [30, 32], but they still rely on scientists to choose the right set of identifiers—a complicated and perhaps impossible task [45, 51]. Moreover, as these frameworks do not constrain *how* data is to be de-identified, specific implementations may be vulnerable to reverse engineering—especially when they try to preserve utility optimally [10], as Chhillar et al. demonstrated on ARX's local recoding algorithm for k-anonymity [9]. Local suppression algorithms have similar issues, one of which we outline in §4.

While differential privacy makes significant strides towards solving these problems, achieving meaningful privacy guarantees can be tricky in practice. Improper handling of tasks such as discretization [16] and floating-point math [26] can compromise differential privacy's guarantee. For example, Stadler et al. discovered that improper handling of metadata by synthetic data algorithms broke privacy guarantees [67]. Additionally, differential privacy is "difficult to reason about" and may offer "little to no protection" when ϵ is high, which may be required to preserve acceptable utility [14].

Finally, systematic disclosure avoidance methods have fundamental trade-offs that may conflict with scientific norms. In an interview study, scientists doubted that systematically de-identified data would remain useful for crucial secondary use cases [21]. In a real-world example, the backlash to differential privacy in the 2020 U.S. Census was driven by concerns over the use of statistical techniques to manipulate that data [4, 61]. Regarding synthetic data in particular, Stadler et al. empirically demonstrated that algorithms can create drastic utility distortions while simultaneously being less predictable and understandable [67], and Kulynych et al. argue that synthetic data is fundamentally inappropriate for many proposed use cases, due to concerns about statistical validity [28].

We see disclosure avoidance tools as still in an early design stage: adoption is low, and many limitations are yet to be compellingly resolved. At the same time, scientists are increasingly expected to de-identify their research data, and they need better support. By investigating the design of disclosure avoidance tools for scientific microdata, our work aims to accelerate and guide the development of better tools that are tailored to scientists' real-world needs.

3 Method

We conducted this user study in person with 23 scientists between August 2025 and February 2026. Participants were randomly divided into two groups; each group used a different tool to de-identify the same real dataset. We interpreted our results via thematic analysis.

3.1 Tool selection

We studied two disclosure avoidance tools that explore different design choices: sdcMicro [73], and SmartNoise [65] combined with SDMetrics [11]. These were selected from a short list of 16 tools that we identified from online advice for practitioners, conversations with domain experts, and related work [8].[3] Our choices are not in any way a recommendation of these tools, given significant usability and privacy issues (see §2.3). However, we believe it is important for participants to anchor their reactions and feedback in experience with real tools and data, which forces them to confront the kinds of trade-offs that necessarily result from systematic disclosure avoidance methods. Below, we explain the design characteristics that led us to choose these two tools.

[3]Our short list included the following tools and algorithms for de-identifying and evaluating data: sdcMicro [73], ARX [57], Amnesia [53], MinBlur [36], the PDPC Anonymisation Tool [56], µ-ARGUS [23], SmartNoise [65], MST [34], PrivBayes [81], DP-WGAN [1], PATE-GAN [27], synthpop [49], MOSTLY AI [38], SDMetrics [11], SynthRO [60], and SDNist [71].

sdcMicro. To explore tools that are *truthful*—that is, tools that reduce detail but do not introduce inaccuracies—half of the participants used sdcMicro [73] v5.7.8 to achieve k-anonymity via the local suppression algorithm, which deletes individual data values. sdcMicro is perhaps the most mature disclosure avoidance tool available, with both an API and a graphical user interface, and it continues to be updated regularly. It is recommended by several de-identification guides [20] and has been used by scientists to protect data published in repositories [21]. In comparison to ARX, another mature disclosure avoidance tool that supports truthful disclosure avoidance, sdcMicro requires less configuration to run its k-anonymity algorithm, and it provides feedback that is more accessible to a lay user (e.g., straightforward visualizations and tables, as opposed to metrics derived from academic literature).

SmartNoise/SDMetrics. To explore differential privacy tools for microdata, half of the participants used SmartNoise [65] v1.0.5 to produce synthetic data using the MST algorithm [34]. As SmartNoise only generates the data, participants used a modified version of SDMetrics [11] v0.23.0 to evaluate the synthetic data. MST won NIST's 2018 Differential Privacy Synthetic Data Challenge and still receives regular updates. Based on an empirical evaluation of synthetic data algorithms [17], MST is a good fit for smaller datasets like the one used in this study (see §3.2). We chose SDMetrics because it is one of the most mature tools of its kind: it is described as a main inspiration for SDNist [63], and it was included by the UK National Health Service in a framework for synthesizing patient data [48]. Compared to other synthetic data evaluation tools that we considered [38, 49, 60, 63], we judged SDMetrics to provide the best balance of being dataset- and algorithm-agnostic, having minimal bugs, and providing interesting metrics and visualizations.

3.2 Dataset selection and preparation

To find a dataset for participants to de-identify, we considered hundreds of candidates from repositories such as ICPSR, ultimately choosing data collected in 2023 for the Detroit Metro Area Communities Study (DMACS) [19]. Using real data improves external validity by capturing realistic risks and characteristics. By using data that has been approved for publication, we can also focus on participants' perspectives on de-identifying data beyond current norms. We obtained permission from the data authors to use the DMACS data for our study, and we adhere to the terms of use, including only sharing data with authorized users and not attempting to re-identify individuals. Participants only had access to the data under our supervision.

The DMACS data comprises survey responses from 1,911 adult residents of Detroit, Michigan.[4] It measures a broad range of opinions and experiences to inform local community development [37]. To make the dataset more manageable for participants, we selected 48 of the original 250 variables, focusing on demographics, health, and finances. Additionally, as the algorithms used in our study are designed for discrete rather than continuous data, we grouped birth year into decades. The Open Science appendix contains the script we used to produce the modified dataset, as well as a codebook and list of survey questions for the modified data.

The modified DMACS data aligns well with the three main criteria we used to pick a dataset:

- **Clear potential for harm.** If individuals are re-identified, they should face serious risks. This data describes respondents' disabilities, debts, evictions, sexual orientation, income, and more. This could influence decisions about health insurance and employment, and respondents may consider some of it to be private or stigmatizing.
- **Re-identifiability.** Individuals should be realistically re-identifiable. This data is constrained to Detroit residents and contains numerous indirect identifiers, including birth decade, gender, race, education, and marital status. In our assessment, there are dozens of outlier individuals who are extremely unique within the dataset and quite likely unique within the total population based solely on traditional demographic identifiers, due to rare characteristics—e.g., one individual who was born between 1920 and 1929, four who are Pacific Islander, and five who are trans women—in combination with other identifiers.
- **Familiar context.** Participants in our study should find it easy to understand the data and to reason about risks and use cases. This data is drawn from the general population of a large U.S. city and focuses on everyday topics such as personal health and household finances.

We note that we do not intend to single out the DMACS data for judgment in any way. This data is not unusual in its level of risk and identifiability, and it is additionally protected by access and use restrictions that may well be adequate for its particular context.

3.3 Task design

Our user study comprised three phases: orientation, tool use, and reflection.[5] These phases are summarized in Figure 1, and protocols are in Appendices A and B. Participants completed tasks in person on a computer we provided, to reduce hassle (participants did not install software) and risk (we did not transmit the DMACS data over the Internet). The recorded portion of study sessions lasted 71 minutes on average.

Orientation. We began each study session by introducing motivations and challenges associated with de-identifying research data, including an example of how individuals can be singled out in a dataset using combinations of indirect identifiers. We introduced the modified DMACS data as a survey measuring 2,000 Detroit residents' opinions and experiences relating to health, economics, and more, with the overall goal of informing local community development. Participants spent a few minutes familiarizing themselves with the data in a spreadsheet viewer. As they explored the data, we prompted participants to think aloud about potential use cases and privacy risks, and then we asked what kind of support they would like if they were to de-identify this data for public release.

[4] For simplicity, we introduced DMACS to participants as if it were a one-time survey; in reality, DMACS is a longitudinal project that regularly collects and publishes data on the same population, which would complicate the task of disclosure avoidance.

[5] We piloted our study within the research team and also conducted three external pilots with scientists, which are not analyzed as part of this study. Based on these pilots, we refined our descriptions of technical concepts and streamlined our study protocol to reduce the number of tool features that participants interacted with.

Figure 1: The three main phases of our study. In the middle phase, participants completed different steps, depending on whether they used sdcMicro (top row) or SmartNoise/SDMetrics (bottom row). This phase was repeated as time permitted.

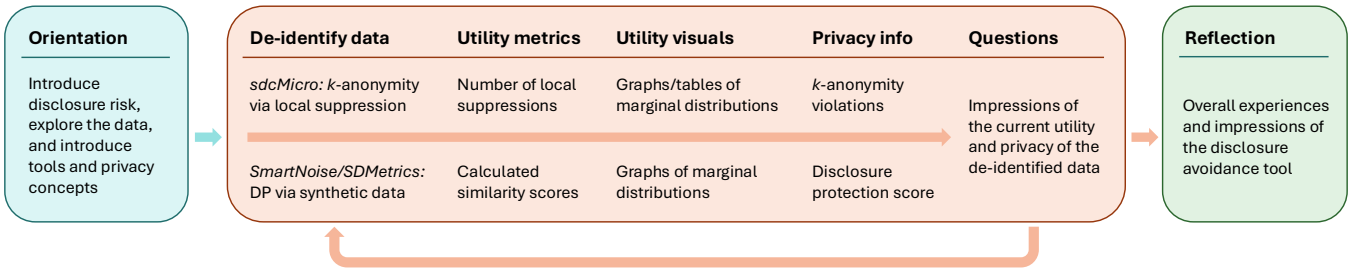


Table 1: This table summarizes visualizations and metrics in each tool that participants interacted with.

Visualization or metric	Description
sdcMicro	
Number of local suppressions	A count of deleted values for each variable; displayed in the table *Information on local suppression*.
Changes in distributions	Comparisons of specific univariate or crosstab distributions before and after de-identification; visualized in graphs (*Barplot/Mosaicplot*) and tables (*Tabulations*) [Figure 2a].
k-anonymity violations	All records violating *k*-anonymity, for a specified *k*; displayed in the table *Observations violating k-anonymity* [Figure 2d].
SmartNoise/SDMetrics	
Calculated similarity scores	Statistical metrics comparing distributions in the original and synthetic data, for each variable and crosstab; displayed, respectively, in the *Column Shapes* and *Column Pair Trends* visualizations [Figure 2e]. Scores are based on total variation distance and range from 0 (non-overlapping distributions) to 1 (identical distributions).
Changes in distributions	Comparisons of specific univariate or crosstab distributions before and after de-identification; visualized in bar graphs (`get_column_plot`) [Figure 2b] and heatmaps (`get_column_pair_plot`) [Figure 2c].
Disclosure protection score	A statistical metric of privacy in a specific attack scenario, where the user specifies identifying variables and sensitive variables. The modeled attacker knows identifying values and tries to guess sensitive values in the original data, based on associations in the synthetic data. Scores range from 0% (all guesses correct) to 100% (no better than random).

Next, we introduced the assigned tool. To reduce bias, we emphasized that we were not involved in its development and that we were seeking honest feedback, positive or negative. For both tools, we described the disclosure avoidance technique and framework that participants would be using. For sdcMicro, we explained how the tool deletes individual values to achieve *k*-anonymity, reducing the risk of individuals being singled out. For SmartNoise/SDMetrics, we explained how the tool samples synthetic data from noisy correlations created using differential privacy, reducing the risk that highly specific correlations could reveal sensitive information about individuals. As it is not straightforward to explain these technical concepts, we developed our explanations iteratively in pilot studies, provided basic visual aids, and invited participants to ask questions.

Tool use. During the main portion of the study, participants followed instructions to de-identify the data and evaluate the resultant utility and privacy. Throughout this process, we offered help proactively when participants were confused, and we prompted them to think aloud about their reactions to the tool and its output. Providing instructions and help enabled us to focus on more fundamental research questions about tool design, by helping participants navigate unfamiliar tools more quickly and with greater understanding. The goal was not to prescribe what disclosure avoidance should ideally look like in practice, but rather to take participants through a reasonable workflow demonstrating how a certain type of tool could be used and what sorts of trade-offs might ensue.

For sdcMicro, we developed a handout with instructions to help participants navigate the user interface. For SmartNoise/SDMetrics, as neither tool has a user interface, we developed a Jupyter notebook for participants to run step by step, without needing to know how to code; we also copied relevant SDMetrics documentation into the notebook, with minor modifications, to provide potentially important context such as the definition of calculated metrics. Both instructional materials are available in the Open Science appendix. Table 1 describes the metrics and visualizations that participants interacted with, and Appendix E contains some illustrative examples.

Participants using sdcMicro reviewed a set of 15 pre-selected demographic identifiers for *k*-anonymity and were allowed to make minor changes before running the local suppression algorithm to achieve *3*-anonymity.[6] We chose a low *k* because this data is a small sample of the much larger Detroit population. Then, participants viewed a table counting deleted values per variable; explored

[6]When pre-processing the DMACS data for sdcMicro, we chose to leave pre-existing missing values as separate categorical values instead of converting them to semantic empty values (i.e., NAs in R). In our own experimentation, applying local suppression with semantic empty values resulted in minimal suppression of sensitive values across the dataset, failing to even suppress any of the four Pacific Islander race values.

graphs and tables conveying changes in user-selected univariate and crosstab distributions; and interacted with a table listing records that would be in violation of different, user-specified k values.

Participants using SmartNoise/SDMetrics generated synthetic data for $\epsilon = 5$, which we chose as a compromise between higher ϵ used for real-world synthetic datasets [22, 74] and lower ϵ that provide stronger privacy guarantees. Then, participants viewed metrics capturing changes across all univariate and crosstab distributions; explored graphs conveying changes in user-selected univariate and crosstab distributions; and calculated a disclosure protection score that aims to measure privacy in a specific attack scenario where the user specifies which variables are identifying or sensitive.[7]

Once participants had made it through the instructions, we asked them to describe how they perceived the utility and privacy of the output data, asking probing questions as necessary. Where applicable, we asked participants to consider the use cases and privacy risks they had articulated themselves during the orientation phase. As a backup, we suggested the use case of researchers including this data in a meta-analysis focused on the drivers of mental health, as well as the privacy threat of a health insurance provider trying to learn individuals' disabilities and finances based on their demographics. Then, time permitting, participants redid the disclosure avoidance process, making different choices. For example, various participants chose a different ϵ, selected different identifiers for k-anonymity, and specified which variables are more important to preserve during local suppression.

Reflection. For the final 15 minutes, we asked semi-structured interview questions about participants' overall experiences and perceptions. Questions covered the usefulness of the tool's information and capabilities, concerns about impacts on secondary use cases, and suggestions for improvements.

3.4 Recruitment

For our in-person user study, we recruited scientists from the authors' two institutions (the University of Maryland and the George Washington University) by contacting faculty in areas such as psychology, epidemiology, and criminology; we invited them to participate and/or share our study with their colleagues and students. We also asked participants to share our recruitment message with other qualified scientists. Prospective participants completed a four-minute questionnaire (see Appendix C) asking about their research area, experience with quantitative human-subjects data, and demographics. From the screening questionnaire responses, we conducted purposive sampling with two criteria:

(1) We prioritized scientists with greater research experience, as they are more qualified to assess how disclosure avoidance methods would impact the usefulness of the data.
(2) We tried to recruit across a diverse range of research areas in the social, behavioral, and health sciences.

All participants had either (1) taken three graduate-level classes in which they analyzed quantitative human-subjects microdata or (2) analyzed such data for two research papers or reports that were submitted to peer-reviewed publications or research funders. Participants were also required to be 18 years or older; speak English; and be a graduate student, faculty member, or staff member at one of the two universities.[8] Participants were compensated $60 in cash.

Table 2: Counts summarizing participants' research backgrounds and experiences with quantitative human-subjects microdata, broken down by which tool they used in our study.

Category	sdcMicro	SmartNoise/SDMetrics
Research area		
Health / medicine	3	4
Crime / justice	3	3
Psychology	3	3
Other social sciences	3	1
Relevant graduate-level courses[*]		
0	1	1
1–2	1	1
3–4	5	6
5+	5	3
Relevant research projects[†]		
0	3	1
1–2	2	4
3–4	4	5
5+	3	1

[*] Number of classes taken in which participants analyzed data.
[†] Number of research papers or reports submitted to a peer-reviewed publication or research funder in which participants analyzed data.

Table 2 summarizes participants' research areas and levels of experience.[9] All participants are PhD students, except for M2 and N1 (faculty members) and N7 (MS student). Appendix D contains more information about research areas and demographics. Participants' aliases are assigned based on which tool they used in the study; an alias prefixed with M indicates a participant who was assigned sdc**M**icro, while a prefix of N indicates Smart**N**oise/SDMetrics.

3.5 Thematic analysis

We conducted a thematic analysis on the data from our user study [5]. The first two authors inductively developed an initial qualitative codebook while reviewing video recordings of four study sessions together. As the high-level codebook structure was stable, the remaining study sessions were coded by the first author, who flagged all points of uncertainty and all changes to the codebook, meeting regularly with the second author to make coding decisions together. The codebook is in the Open Science appendix. To facilitate coding, we took audio and screen recordings; after generating transcripts,[10] we listened through each recording to fix transcription mistakes and annotate transcripts with participants' actions, such as which variables they chose to visualize using their tool.

[7] We made minor changes to SDMetrics to fix some distracting usability issues. See the Open Science appendix for a description of these changes and a link to the code.

[8] We checked institutional affiliation as a straightforward way to comply with the DMACS dataset's terms of use.

[9] This information comes directly from screening questionnaire responses, except for the counts of participants who have de-identified data for publication, which we evaluated conservatively based on participants' qualitative descriptions. This is because some participants interpreted the relevant screening question over-broadly.

[10] We used Whisper [52] for transcription and pyannote [58] for speaker diarization. The code was run entirely locally, and no data was shared with third parties.

Our codebook focuses on concepts related to our research questions. To develop themes, the research team discussed overarching patterns that connected multiple participants' perspectives. We mapped each theme to a set of codes, allowing for a single code to associate with multiple themes. To review and refine themes, we read through the coded segments for each theme and wrote a few sentences describing how the codes construct the theme.

3.6 Limitations

Participants could have been biased toward positive feedback if they associated us with the tools and methods studied. Therefore, we emphasized that the tool participants used was created by other researchers, and that we were looking for honest feedback that could identify issues from scientists' perspectives.

Participants' opinions on tools may have been affected by usability issues, which is outside the main scope of our study. To reduce usability issues, we provided detailed usage instructions and offered help proactively when participants were confused. Providing this guidance, which would not be available to scientists outside of our study, may have affected external validity. However, we believe that real-world adoption of disclosure avoidance tools by scientists will likely require similar forms of guidance, for example in the form of community-developed instructions and norms.

Our study has a between-subjects design: each participant interacted with just one of the two disclosure avoidance tools. We made this decision to keep the duration and cognitive load manageable for participants, but it limits our ability to directly compare tools.

We recruited participants from only two institutions, which may limit generalizability; we judged this to be worth the benefits of interacting with participants in person. We did not observe evidence of academic monoculture, as both institutions are large research universities, and participants came from 14 different departments.

Finally, this study explores how individual scientists respond initially to disclosure avoidance tools; in reality, norms are formed over time within scientific communities. Our findings provide novel insights into the design of such tools, but they should not be taken as predictions of how scientific communities would respond, should such tools receive more widespread attention.

4 Context on disclosure avoidance outcomes

To provide context for interpreting participants' interactions, we first provide a brief overview of how sdcMicro and SmartNoise tended to affect the DMACS dataset when the instructions we provided were followed. Participants observed similar, but not necessarily identical, impacts to the ones described below: SmartNoise is non-deterministic, and while sdcMicro's local suppression algorithm is deterministic, participants often made slight changes to the set of identifiers used to calculate k-anonymity.

sdcMicro deleted individual values in the data (*local suppression*) to achieve *3*-anonymity for a set of 15 pre-selected demographic identifiers. Birth year was most frequently suppressed, with around 1,140 out of 1,911 values deleted. Around 390 household income values were suppressed, along with 230 for gender, 130 for marital status and education, 60 for both sexual orientation and Middle Eastern ethnicity, and fewer than 40 for all others, mostly comprising binary race variables (i.e., a separate yes/no variable for each individual race). The pattern of suppressions tended to preserve the directionality of trends—for any given variable, prevalent categories remained prevalent in the de-identified data—but the shapes of distributions were affected, with rarer values being disproportionately removed. The local suppression was notably vulnerable to being reverse engineered: e.g., for the Pacific Islander race variable, sdcMicro suppressed all four *Yes* values and often only those four (sometimes one or two more), enabling an attacker to conclude with confidence that all (or most) of the removed values were *Yes*.

We included sdcMicro in our study not to investigate participants' perceptions of these specific trade-offs, but rather, at a high level, how they engaged with the kinds of impacts it has: deleting values and distorting distributions, based on a selected set of identifiers, while preserving truthfulness in the data.

SmartNoise generated a differentially private synthetic dataset with ϵ of 5 using the MST algorithm, which automatically detects correlations in the original data, adds noise, and then samples synthetic records. Distributions of individual variables were largely preserved across the board. SDMetrics uses total variation distance to measure the change in distributions from the original data; the greatest change for any variable was around 0.03 (out of a maximum 1), and the proportion of any given category tended to be within 1% of its original proportion. Distributions of crosstabs were more affected; some had a total variation distance of as much as 0.2. Crosstabs involving household income were consistently among the most impacted; trends were preserved but noticeably noisier (e.g., higher education levels were overall correlated with higher household incomes, as they were in the original data, but the gradient was much less smooth). For some highly correlated crosstabs, such as one type of disability with another, SmartNoise appeared to lose some of the strength of those correlations, altering the proportion of some combinations of values by 5% or more.

We included SmartNoise/SDMetrics not to investigate these specific tradeoffs, but rather participants' high-level engagement with its inherent characteristics: preserving correlations by design (especially with high values of ϵ), de-identifying data agnostic of threat model, and replacing truthful data with synthetic data.

5 Findings

Our findings are organized by research question: first perspectives on tool acceptability, then interpretation of metrics and visualizations, and finally perceived gaps and strengths in tool workflows. Each bolded header represents a theme from our analysis.

5.1 Acceptability of disclosure avoidance tools

To begin, we consider how participants perceived the utility and privacy of the de-identified data. We find that significant variation in how different participants perceived the utility and privacy of essentially the same de-identified data led to vastly different conclusions about the overall acceptability of disclosure avoidance tools. Participants who were doubtful of the identifiability of the original data, or were taken aback by the scale and kinds of changes made to the data, wound up deeply skeptical about the usefulness of these tools. On the other hand, participants who saw a need for increased privacy, and bought into tools' metrics and visualizations, spoke cautiously but optimistically about tools' potential.

Skepticism about baseline risk causes both disclosure avoidance methods to be seen as extreme. Participants had differing opinions on the usefulness of systematic disclosure avoidance tools and the utility of the data they produce. These differences appeared to be driven more by discrepancies in how participants perceived baseline disclosure risk than by any other factor that we could observe. This is because participants' perception of the baseline risk informed whether the substantial trade-offs of both tools were seen as necessary and something scientists must therefore work around, or whether they were seen as nonsensical and extreme.

A substantial portion of participants did not believe that the original dataset was realistically identifiable, giving two main reasons. First, participants argued that individuals would not be identifiable within the much larger population from which the DMACS data is sampled. As M4 said, for a health insurance provider to re-identify a data subject "would be a little delusional, because it's still a random sample of 2,000 people in the Detroit area." They elaborated that an individual who is distinctive within the dataset—e.g., someone "in their 40s or 50s that has a unique set of disabilities"—is unlikely to be unique within Detroit. Second, participants argued that the data lacked sufficient detail to identify individuals. M2 pointed to a lack of granularity: "I think it's helpful here [for privacy] that you have ranges instead of specifics, especially with birth year. . . . Anything related to health or a personal diagnosis is a dichotomous variable, so nothing is really personal there." Others noted that some data may be somewhat unreliable or obscure: M1 argued that, as self-reported data, household income "might not be accurate, and you might not be able to know exactly what the person's household income is, based on available data or even if you personally know that person." While we as the research team acknowledge that these factors do indeed reduce disclosure risk, we nevertheless believe that some individuals in the dataset are likely identifiable (see §3.2).

For several participants who felt the most strongly that the tool they used is not a reasonable approach to dislosure avoidance, skepticism about risk played a key role in leading them to that conclusion. For example, M1 said, "Considering this is a small sample of the residents from a large city, I think [sdcMicro] has been too conservative. . . . We now have 16% of missing data on marital status. And my question would be, is that necessary? Even if I know a person who is married or not married, combined with all of the other information, will I be able to locate that person in the population?" Similarly, citing extreme differences in distributions between the original and synthetic data, N2 concluded, "I feel [SmartNoise/SDMetrics] is taking a pretty overly conservative approach on a dataset that isn't super granular to begin with."

To achieve acceptable utility, participants sometimes tried to significantly reduce privacy. Several sdcMicro users reduced the set of variables they considered identifying, which fundamentally shrinks the scope of protection provided by k-anonymity. For example, after seeing the number of race values suppressed during the first go-around, even though they initially pointed out the identifying potential of an individual's race, M12 decided not to include race variables as identifiers the second time. They justified this decision retroactively by arguing that race is crucial for a lot of sociology research and not so identifying in a diverse community: "It's an immigrant society in America; you have lots of people from different races." Similarly, though M11 initially decided to include political affiliation as an identifier, seeing the number of suppressed values overall led them to reconsider whether it and other variables such as household income are actually identifying: "Am I over-including variables that make the k score not very useful?" In contrast, SmartNoise provides users with less control over the disclosure avoidance process; some participants tried increasing ϵ (often to 10, as we told them when introducing differential privacy that real-world synthetic datasets have used ϵ of about 10), which generally resulted in marginal improvements to perceptions of utility.

Participants were hesitant to rely on technical disclosure avoidance methods. The changes that tools made to the data were seen as too extreme for many reasons, with concerns about decreased utility first and foremost. Commenting on the number of values sdcMicro deleted, M7 said, "If I'm the user of this data after its publication, I would be pretty upset. . . . It lost so much statistical power." Observing changes in specific crosstab distributions, N8 worried about the impact on statistical analyses: "If the effect size for household income and birth year would be 0.2, but now the effect size becomes 0.4, then I would say that's not okay."

Beyond data utility, some participants also disliked the idea of relying on tools, which can feel opaque in design. M11 said, "I'm the kind of person who wants to know the ins and outs of all of my data. I don't want to just press a button and trust the algorithm that it treated it totally fair and randomly." Similarly, N8 felt nervous about trusting code that they had not written themselves, saying, "I'm not an engineering major or a CS major, so I don't really understand the code or anything behind this"; they noted that wider adoption of a tool by other scientists would be a reassuring sign.

Some participants also observed that systematic disclosure avoidance methods can clash with norms within their scientific communities. This was mainly a concern with SmartNoise; participants anticipated that reviewers, advisers, and other scientists would automatically reject synthetic data. For example, N1 said that in the discipline of criminology and criminal justice, if one were to publish a synthetic version of a survey dataset, "Whether it's appropriate or not, there'd be red flags. I think it's a cultural thing." After learning that SmartNoise distorts marginal distributions to achieve differential privacy, N8 said, "As epidemiologists, we want to see the true correlations. If you are manipulating the correlations, why should we even report that if we know that's fake?"

As a result, some participants expressed a preference to rely less on technical or systematic methods for protecting data privacy. M11 argued that, for sensitive data, restricting access and use to scientists who have shown they can be trusted to handle the data properly is a better approach than local suppression: "I just think also you have to think about not always the most technical way to make things least identifiable."

Trying out tools firsthand can increase perceptions of acceptability. Interacting with tools increased some participants' confidence that de-identified data would be useful for at least some secondary use cases. In particular, it was not always intuitive that it was possible to produce data with similar marginal distributions. N6 said, "When you first started talking about using data that's synthetic, I was like, I don't know about that. But I feel a lot more positive toward it than I thought I would have," explaining that

interacting with SDMetrics's utility metrics and visualizations increased their belief that the synthetic data could produce similar research findings. Similarly, after examining graphs generated by sdcMicro comparing univariate distributions before and after local suppression, M6 said, "I would have expected . . . a different-looking distribution for some of those variables where so much got deleted from. However, it looked visually very much similar."

Participants who saw significant benefits to using tools, especially improving privacy, were often more optimistic about their usefulness overall. Participants who perceived higher levels of baseline disclosure risk tended to point to the presence of rare values and data types that could be commonly known or publicly discoverable; these participants often felt that impacts on utility were more acceptable. For example, M10 felt that many of the identifiers in the dataset could be linked to public sources such as person finder websites, LinkedIn, and government records; towards the end of the study session, they reflected, "I think it's very easy for quantitative researchers to lose the ball that these statistics and numbers are actual people with actual lives. . . . I think this tool is very useful for the interest of protecting participants." Some also hoped that tools could enable publication of more data: M5 said of sdcMicro, "In different labs and research programs that I've worked in before, the de-identification process is essentially just deleting entire columns of data, versus an approach like this, which I think preserves a lot more of the data."

5.2 Interpreting data metrics and visualizations

We now take a closer look at how participants interacted with tools to make sense of the utility and privacy of the de-identified data. As noted previously, our goal is not to evaluate the specific details of any given interface, but rather to understand at a high level what *kinds* of interfaces could be useful to scientists and why. We find that, by combining high-level metrics and low-level visualizations, tools have made a promising (though limited) start to help scientists evaluate data utility in intuitive and constructive ways. However, features for evaluating privacy were less beneficial, indicating a need for designing better functionalities or interfaces.

Combining metrics and visualizations formed a positive feedback loop for understanding data utility. To evaluate utility, both tools incorporate high-level metrics and low-level visualizations (see Table 1). Individually, both types of elements face significant barriers to understanding: e.g., the metrics may be overly abstract or opaque, while the visualizations provide too many options for users to take them all in. However, participants generally converged on a similar process that combined metrics and visualizations iteratively to build up their understanding of the de-identified dataset's utility. We describe that positive feedback loop below.

First, participants viewed a summary of high-level metrics—counts of deleted values for sdcMicro, and calculated similarity scores for SDMetrics—across multiple variables. Though some expressed uncertainty about how to interpret these metrics, participants were often comfortable using their intuition to draw preliminary conclusions about utility, especially when metrics were particularly high or low. For example, looking at similarity scores in SDMetrics, N7 said, "I don't know the standard, but if something is 97%, well, it's high." Similarly, most participants quickly decided that the percentage of birth-year values suppressed by sdcMicro (typically 60%) was probably concerning, and some suggested heuristics for assessing other variables. For example, M3 said, "Surely there has been a time in a statistics class when I've learned, like, 'More than this amount missing is problematic.' The number I have in my head is more than 5%." Engaging with these metrics caused a few participants to raise concern about impacts on certain demographic variables that are important for future use cases broadly: M1 said they would not want gender to be suppressed at all, explaining, "Gender is a basic control in all of the social science studies. . . . I would want to keep all the age, gender, race, ethnicity information."

Next, participants prioritized specific variables and crosstabs to visualize at a low level, by generating graphs and tables of marginal distributions before and after de-identification. Given the large number of potential visualizations, participants often relied on the metrics to focus their attention on the variables and crosstabs that were most impacted: e.g., N6 selected crosstabs involving disabilities, education, and household income to visualize specifically because they had lower similarity scores. Using these low-level visualizations, participants tried to understand how distributions had changed: e.g., M12 examined tables comparing distributions for each race variable, concluding that too high a percentage of values had been suppressed for Pacific Islander (4 out of 4 *Yes* values deleted) and Native American (32 out of 37 *Yes* values) races to remain useful, but that utility was acceptable for Black (12 out of 1,303), White (19 out of 359), and Asian (14 out of 38) races.

As they examined visualizations, participants refined their understanding of metrics, forming positive feedback loops. For example, based on counts of local suppressions, M6 picked variables with many deleted values to inspect graphically; based on the graphs, they concluded that the high suppression counts were not as bad as they could be, because even for birth year (the most heavily impacted variable), "the overall distribution still looks very similar." N5 reflected that interacting with the "more straightforward" crosstab graphs was helpful while they were still getting a handle on the similarity scores, as "it took a while to just really get understanding about what the score means or how to interpret those scores."

Furthermore, some participants were able to follow this process to compare different versions of the de-identified data. For example, while experimenting with different values of ϵ, N4 observed that the combination of household income and education consistently had one of the lowest similarity scores. As a result, they visualized that specific crosstab as they tried out different values of ϵ. At $\epsilon = 10$, N4 noted that the distributions "look more similar" than with $\epsilon = 5$, while also pointing out some remaining differences; this informed their overall conclusion that the dataset's utility was "definitely better, but marginally, just because it already looked pretty good." In contrast, at $\epsilon = 1$, N4 expressed strong concerns about the data's usefulness, as distributions appeared to be "flattened out."

However, existing utility metrics and visualizations obscure important nuances. Some features, such as SDMetrics's display of crosstab similarity scores, were so visually and informationally overwhelming that some participants actively avoided them. M5 expressed a common reaction to sdcMicro's graphs of crosstab distributions: "Oh god, I can't read that. I just can't, there's too much

going on here. I'm gonna have to reset myself." SDMetrics's graphs of crosstab distributions, which use color rather than size to convey the frequency of different values, were generally better received, with several participants using them extensively and speaking positively about their design. However, the visual comparison they demand can still be challenging: N4 wished for an interface that would show more explicitly the difference in frequency between the original and synthetic data for corresponding cells in the graphs. Making differences between marginal distributions in the original and de-identified data easier to notice and quantify was an extremely common piece of feedback across both tools.

Metrics and visualizations can also cause users to overlook or misinterpret important nuances. We observed some participants making unsupported statements about complex visualizations: e.g., when looking at SDMetrics's graph of household income distributions, which compared the frequencies of 15 different categorical values, N9 observed incorrectly that the frequency of the middle income categories had changed less compared to categories at either end. Simple metrics can also obscure important details: sdcMicro's local suppression table shows just the quantity of deletions for each variable, which can obscure outsize impacts on certain types of variables. For example, even though a very small proportion of all values (1–2%) was typically deleted for the Native American race variable, since that included most of the *Yes* values, most of the utility for understanding Native American experiences was gone. Participants rarely noted these kinds of nuances, instead often concluding that low overall proportions of suppressed values indicated good utility. Inspecting graphs and tables sometimes helped users spot these sorts of issues, but relying on users to select the right variables and notice outsize impacts is not ideal.

Participants using both tools expressed that more descriptive statistics and other calculated values would help them move beyond visual comparisons and attain a deeper understanding of data utility. For instance, N3 said, "If you're using a linear regression, it assumes you're normally distributed. But if it's no longer normally distributed, then assumptions could be violated. . . . You ideally would want more context about *how* did the distribution change, not just how much did it change." Several participants using sdcMicro wanted to know if values were missing at random: M1 said, "If all the missing [values] are missing at random, it also means that when the users use a dataset, they can either keep all the complete observations with non-missing values, or they can impute missing data using different strategies—it won't bias the results." In reality, suppressed values are unlikely to be missing at random, as they were suppressed due to attributes within the dataset, but these comments indicate more generally that participants using sdcMicro wanted better methods to understand properties of missing values as pertains to secondary use cases.

Finally, participants largely accepted these tools' limited framing of utility as the preservation of marginal distributions; few spoke in any detail about impacts on more complex, multivariate relationships.[11] To some extent, this shallower lens makes sense, as there is usually some uncertainty about future use cases in open science, which can make it harder to decide which of many possible multivariate relationships to prioritize. How to address this gap is an open question. N9 envisioned SDMetrics enabling users to choose variables for a regression that would be run on both the original and the synthetic data, in order to compare p-values, coefficients' confidence intervals, and other results. However, we note that the benefit of testing user-specified analyses on the de-identified data is not entirely clear: if a desired analysis is known in advance, then scientists could simply release the outcome of that analysis on the original data. One option is to have the tool generate a variety of potential secondary analyses to test, but, when asked, N9 felt that letting users choose would be more useful. We leave this question for future work to explore.

Features for evaluating privacy were largely uninformative. Participants struggled to gain useful information about privacy from both tools. sdcMicro interfaces that could be used to evaluate privacy—such as visualizations of marginal distributions and k-anonymity violations—were largely used to evaluate utility instead.[12] For example, participants primarily used the table of k-anonymity violations to compare the utility trade-offs of choosing higher values of k, rather than the state of existing risks. As M2 said, "You could use that to tease out the amount of data you're willing to potentially lose. . . . I might play with this and be like, okay, maybe we should have k equal 6. Then I'm like, ooh, 1,100 observations [violating *6*-anonymity]—that's a lot, so maybe we settle for *5*." We believe this is because participants generally lacked stances on appropriate values of k, so the number of violations for any given k was not meaningful to them from a privacy perspective.

We also note that participants generally did not notice the privacy vulnerabilities in sdcMicro's disclosure avoidance implementation. Only one participant, M12, pointed out that by suppressing only five Pacific Islander race values, sdcMicro largely failed to hide the four *Yes* values: "That's weird, you suddenly have five NA for Pacific Islanders—it means they are Pacific Islander!"

Aside from ϵ, which is famously difficult to reason about [14, 15, 40, 43, 64], the main information about privacy provided to participants using SmartNoise/SDMetrics was the disclosure protection score, which measures risk in a specific attack scenario, based on user input on which variables are known to an attacker and which are sensitive. As with SDMetrics's similarity scores, disclosure protection scores towards the middle are harder to interpret than scores at either extreme.[13] Unfortunately, while tools provided a positive feedback loop for understanding utility, they lacked salient low-level visualizations to help participants make sense of the disclosure protection score. As a result, participants often drew conclusions about middle-range disclosure protection scores with questionable justifications: N7 suggested that 65% was not good because it was

[11] It is possible that some participants withheld their opinions about multivariate analyses because they considered tools' impacts on simple properties of the dataset to already be disqualifying. Still, very few participants—positive or negative about tools—spoke in detail about tools' impacts on more multivariate relationships.

[12] We considered having participants interact with other sdcMicro privacy metrics, such as a predicted number of re-identifications (listed in the tool interface under *Risk measures for categorical key variables*), but pilot participants found these metrics particularly confusing, and they failed to generate interactions worth studying further.

[13] Additionally, we would argue that the disclosure protection score's awkward definition exacerbates this challenge. For one, it is calculated relative to a naïve attacker who guesses sensitive values completely randomly, regardless of actual distributions, meaning that the implications of the same number can differ dramatically depending on which variables were picked. The score also tends to increase dramatically as the number of sensitive variables grows.

too close to 50%, which is halfway, while N3 said scores below 90% or 95% would be concerning, explaining, "Maybe that's me working off of confidence intervals, that I want a 0.05."

Additionally, though disclosure protection scores only measure privacy under specific attack scenarios, participants often extrapolated from a single score (or a small set) to the overall level of privacy. For example, based on just five scores around 65%, each with political affiliation as the sensitive variable, N5 concluded that overall privacy was good. While providing scientists with the capability to assess specific attack scenarios is useful, we believe tools should work to ensure they do so with some comprehensiveness.

5.3 Gaps and strengths in tool workflows

Finally, we analyze how disclosure avoidance tools could be improved to meet scientists' needs in terms of overall workflows. Based on what participants experienced as gaps and strengths of the tools they used, we characterize two main potential improvements: providing better support for decision-making, and generating documentation to inform secondary use. We also describe differing perspectives on whether tools should give scientists greater control over how disclosure avoidance algorithms are applied.

Participants wanted better support for decisions involving setup and evaluation. Perceived gaps were often related to helping scientists evaluate the utility and privacy of the de-identified data. For example, N4 said that guidance on acceptable thresholds for metrics such as SDMetrics's similarity scores "would be helpful for the average researcher who doesn't really know how to do these things. . . . I just have a lot of questions about how this works and how to evaluate it." Participants noted this kind of guidance should be based on discussion within a broader scientific community: N6 expressed a need for "other people to help consider what our metrics are that we would use to say we are confident that we can use the synthetic data." Guidance on how to interpret metrics could improve the currently limited usefulness of metrics and visualizations for evaluating privacy in particular (see §5.2).

Other gaps were related to helping scientists de-identify the data to begin with, including setup and parameterization. For example, multiple participants suggested that, based on the topic and other characteristics of the dataset, tools could provide advice on appropriate values of ϵ, variables to prioritize, and more. Several participants saw sdcMicro's table of records violating k-anonymity as a particularly good example of a responsive interface to assist with decision support. As mentioned in §5.2, many participants liked the idea of using a similar interface to rapidly evaluate utility trade-offs of different disclosure avoidance configurations, such as different values of k, without de-identifying the data anew each time. Several participants further described hypothetical improved versions that would provide more specific information about disclosure avoidance impacts: M3 said, "This [currently] just shows you all of the observations. I would want to know, easily, if I set k to 5 instead of 3, what percentage of birth year would be suppressed?"

Participants wanted tools to generate documentation to inform future use of de-identified data. Several participants indicated that using de-identified data for research would require additional documentation: e.g., after viewing SDMetrics's utility metrics and visualizations, N4 noted that it would be important for scientists analyzing synthetic data to have those same statistics to understand the potential degree of inaccuracy in their analysis. This was seen as particularly important because disclosure avoidance tools may have different impacts on data utility, depending on use case. N9 noted that the scientists de-identifying the data might have different priorities than secondary users, which should be made clear: "I could be very concerned about health and disability, but not that concerned about the financial situation. . . . That's why I feel the communication between the people who are de-identifying and the people who will be using the data at the end matters."

Providing this important documentation would need to be weighed against its potential to increase disclosure risk, which M2 acknowledged, suggesting a summary of values suppressed by sdcMicro, including blurred statistics ("a confidence interval . . . like the real number of males is between X and Y; we censored it, so we can't tell you exactly, but it's gonna be within this threshold") and formal concepts ("whether or not the missingness is random"). We note that differential privacy is a good fit for this particular goal, as information about the algorithm and parameters including ϵ can be published without breaking the privacy guarantee; depending on the algorithm, this could enable the safe publication of information such as confidence intervals for specific marginal distributions.

There is a lack of consensus over whether tools should give scientists greater control over how disclosure avoidance algorithms are applied. Several participants suggested that users should have more control over nuances of disclosure avoidance algorithms. In particular, many wanted to mark certain variables as priorities for preserving data utility. N9 suggested that SmartNoise could support grouping variables into categories, such as demographics, health, and finances, and then picking different values of ϵ for each category: "If I want my synthetic data to be very close to my original data in terms of the demographics, that one can have a higher ϵ and then the other areas can have a lower ϵ." Prioritizing individual identifiers is a capability that sdcMicro supports, which some participants interacted with on their second try. While participants liked the idea, they felt that the design of the feature needed improvement: nearly every participant who tried it commented on how odd it was that sdcMicro forces users to give each identifier its own distinct rank, rather than allow users to just specify a subset of identifiers they do (or do not) prioritize. M6 just wanted to mark birth year as a high priority but did not like having to prioritize different race variables over one another: "It feels a little silly to be like, this race gets a higher priority than this."

However, some participants also expressed the opinion that greater control is not necessarily better. M3 experimented with prioritizing variables, before concluding that it might be best to run sdcMicro in a more default manner to mitigate users' individual biases: "I thought gender was really important and birth year was not that important to keep, but somebody else might think differently." Similarly, N10 worried that having individual scientists choose ϵ would be "too subjective," suggesting that scientists should instead follow suggested values for different contexts, such as a higher value for general population-level surveys and a lower value for sensitive data from a hospital. Perhaps worse,

scientists could control disclosure avoidance algorithms to artificially support their preferred findings: N11 noted that "sometimes as researchers, we are so inclined in having certain good outcomes to show, that we may be seduced by the certain outcomes of the synthetic data—for example, where we want to show that we have discovered something." As these several participants pointed out, empowering individual users does not necessarily lead to the best outcomes for a scientific community.

6 Discussion

Disclosure avoidance tools could empower scientists and protect research participants' privacy, but they will be most impactful if they are tailored to scientists' needs and concerns. Here we discuss recommendations for tool design drawn from our findings.

Helping scientists understand disclosure risk is a critical need for tools to fill. Systematic disclosure avoidance methods have serious drawbacks—if not the specific drawbacks of the two tools in our study, then other drawbacks. As a result, tools were often a non-starter for participants who were skeptical about the baseline disclosure risk. We believe that disclosure avoidance tools for all classes of practitioners should improve their communication of disclosure risk—especially tools intended for scientists to use. As we find, however, this is challenging, with participants often finding information irrelevant or misinterpreting it. Relatedly, Nanayakkara et al. noted challenges with a percentage-based disclosure risk visualization they designed, as health data practitioners were skeptical that a single numerical score would be useful for making sense of risk in practice [40]. While demonstrating actual re-identification attacks on a dataset could be particularly compelling, this may not be feasible for tools to do on their own: the most convincing attacks from the research literature involve individualized, expert judgment to find external data and devise identity linkage methods [10, 44, 68, 70].

To help scientists understand disclosure risk, tools must accomplish two interrelated tasks: communicating the disclosure risk of the original data, and showing how that risk did or did not change after de-identification. Facilitating better understanding of relevant privacy parameters would be a good start, but tools should think broadly about what information is important in the first place, as well as how it relates to the threats and use cases relevant to open-science data. For example, k can be useful as a relatively intuitive heuristic, but users must also be informed about potential reverse-engineering vulnerabilities (if applicable), so as to avoid giving the false impression that k-anonymity alone can be trusted to totally prevent re-identification. Similarly, ϵ quantifies guaranteed limits on disclosure risk, but the meaningfulness of those guarantees is practically nonexistent at high values, certainly including the values around 10 used for real-world synthetic datasets [22, 74].

There are some promising efforts to develop new frameworks that define disclosure risk in ways that may be more aligned with the context of open science; this could help to overcome perceptions of systematic disclosure avoidance methods as extreme. For instance, Bun et al. propose a formal privacy framework with a weaker threat model than differential privacy, which could entail smaller utility costs while simultaneously being easier to reason about in relation to real-world attacks [6]. Other examples include MinBlur, a k-anonymity variant that increases utility by estimating k within the total population from which a sample is drawn [36], and SafePub, a differential privacy algorithm that preserves truthfulness (i.e., does not introduce inaccuracies) [2]. Further user studies should be conducted to understand how tools incorporating these frameworks affect scientists' perceptions of disclosure risk.

In our study, participants strengthened their understanding of data utility through positive feedback loops that combined high-level metrics with low-level examples (see §5.2); similar strategies could enable tools to convey disclosure risk more effectively. We propose that tools could show users the records that are estimated to be the most identifiable or sensitive, in order to create positive feedback loops that strengthen users' understanding of data privacy. For example, a tool like sdcMicro—where users suppress or coarsen a set of identifying variables—could display the records with the rarest combinations of those values, in order to ensure that users take into account the most vulnerable individuals and how their risk level changes with different de-identification choices. With appropriate privacy safeguards, AI could help pick these records, for instance using an approach like Zheng et al.'s technique for estimating an individual's global k value [82]. Similarly, a tool like SmartNoise/SDMetrics—where differential privacy is applied to generate synthetic data—could show users the original records that are at highest risk of inference attacks, for different values of ϵ.

To further support productive conversations and establish norms, we believe that meta-infrastructure like the proposed public registry of differential privacy deployments [14, 41] would be beneficial. There is currently a great deal of momentum focused on supporting differential privacy practitioners, including the differential privacy registry, but much less energy has been put into helping scientists with disclosure avoidance more broadly. A registry for scientists could document information about disclosure risk and about disclosure avoidance methods and outcomes, in a format that facilitates comparison; this could help scientists understand, for example, how the disclosure risk of their dataset compares with other published datasets. Devising appropriate metrics and evaluating the registry's format is an open direction for future work at the intersection of technical privacy and practitioner-centered research.

Improving guidance should also be part of the solution: our prior work found that disclosure avoidance guides have notable gaps that may mislead scientists about risks, and they fail to provide good instructional examples to show scientists how to assess disclosure risk [20]. At a minimum, guides should provide a step-by-step example of singling out an individual, so that scientists understand the basic attack that they are tasked with preventing. Ideally, guides would be interactive—for example, a webpage with executable code blocks—to give scientists a more tangible sense of the process and level of effort required to re-identify data. Ultimately, as disclosure avoidance is a secondary concern for scientists, it is essential that improved guidance, tools, and other support be embedded in scientific infrastructure, in places where scientists will have access to them exactly when they are needed. When there are considerations that scientists need to plan ahead for, they should be provided with reminders and accountability—for example when applying for funding, getting research ethics review, and especially submitting data to repositories and funding agencies.

Tools should support iterative decision-making, but this also creates new issues that must be accounted for. Understanding the impact of disclosure avoidance on utility is a complex task, and positive feedback loops between high-level metrics and low-level visualizations can help. The tools in this study have made a good start, but future disclosure avoidance tools for scientists will ideally incorporate utility metrics that measure more than just the magnitude of change to marginal distributions; to the extent that those high-level metrics can be paired with low-level visualizations and examples to make them more concrete, users will benefit.

We were particularly surprised by the strength and consistency of participants' reactions to sdcMicro's table of k-anonymity violations—that an interface like this would be invaluable for rapidly assessing the potential impacts of different setup and parameterization choices. We believe this indicates an unmet need for tools to support more iterative decision-making. As it is, most workflows in sdcMicro and SmartNoise/SDMetrics are not very iterative: for one, trying out different values of k or ϵ requires users to restart from the beginning, potentially waiting for several minutes before they can see how outcomes have changed.

Tools should strive to make iterative interactions easier and faster. ViP, which visualizes privacy-utility trade-offs when releasing differentially private statistics, provides a good preliminary example of how disclosure avoidance interfaces for practitioners could be more iterative and responsive; when the user changes ϵ, a dashboard updates dynamically to show a range of possible outcomes across multiple variables in the dataset [40]. Designing analogous interfaces for de-identifying microdata is an interesting and likely even more complex topic for future work.

At the same time, empowering the human in the loop creates new issues that tools should account for. sdcMicro has users choose which variables are identifiers; in our study, some participants backtracked on their initial assessments after seeing the impact on data utility, questioning whether the set of identifiers needed to be so large. This is not inherently a bad thing—disclosure avoidance requires compromises between privacy and utility—but it does show how tools like sdcMicro that give users more agency (as opposed to SmartNoise, for example, which operates with minimal user input) also give users more ways to cut back on privacy, creating risks. Participants also explicitly worried about whether giving users greater control would result in overly subjective decision-making and even enable skewing data towards desired research findings.

Creating community norms is essential to resolve these issues, for which there is no objectively ideal solution. Based on their needs and values, scientific communities could set expectations for disclosure avoidance methods, parameter values, ways of measuring utility and privacy, and thresholds for utility and privacy metrics. Ideally, good norms would leave room for scientists to de-identify data in accordance with the circumstances of their particular research project and with their own personal values, while also simplifying disclosure avoidance and smoothing out individual idiosyncrasies that complicate secondary usage of the data. Once again, meta-infrastructure akin to the proposed public registry of differential privacy deployments [14, 41] could help to create norms; research funders and data repositories could also set their own expectations. Tools could help by integrating guidance from specific scientific communities into tool interfaces and workflows.

Tools should provide better interfaces to help users interpret information. The format of information provided by tools can lead to misinterpretations and oversights, as users may struggle with tasks such as visually comparing graphs or recognizing edge cases for a utility metric. Some of this can be fixed by addressing basic usability issues. While this is not within the scope of our study, we note one simple usability-adjacent suggestion that came up frequently: in addition to showing users graphs of marginal distributions before and after disclosure avoidance, also provide one more graph that specifically plots the difference between the two distributions, to make comparison as easy as possible. Tools could also bring potentially noteworthy changes directly to users' attention, for example by highlighting variables whose marginal distributions have changed significantly as a result of deleted values.

Better visual interfaces could help by enabling scientists to quickly understand changes across a dataset at a high level. Researchers may be able to take inspiration from existing work in related domains of practitioner-centered interface design: for example, an interface designed by Munechika et al. to help machine learning practitioners audit their models for bias could be adapted to help scientists understand changes in crosstab distributions [39].

Finally, AI could provide individualized interpretive assistance, summarizing key takeaways and highlighting changes that users may not notice. Ensuring that AI answers can be trusted is crucial; at the least, they should be grounded in pre-defined considerations related to a specific disclosure avoidance method, and they should be transparent about how they were generated. The problem of trustworthiness may not be fully solvable, so tools should ensure that AI guidance is accompanied by other forms of guidance drawn from sources such as human experts and community norms.

Tools need to provide transparency about disclosure avoidance methods and impacts without compromising privacy. Echoing participants in a user study by Sarathy et al., participants in our study expressed concerns about trusting tools that feel like black boxes to de-identify data correctly and effectively. Participants also noted the need to help secondary data users understand how the de-identified data differed from the original. To address both of these concerns, tools should provide clear summaries and other documentation that can help the scientists de-identifying the data, as well as secondary data users, verify that the data has been transformed in expected ways.

As noted in §5.3, though, these important transparency mechanisms need to be carefully balanced with privacy: sharing information such as disclosure avoidance steps, statistics about the original data, and details about record-level changes can sometimes undermine the privacy protections that tools are meant to provide. Due to its formal definition, which leads to properties such as immunity to post-processing, differential privacy makes it possible to unambiguously determine what additional information can be shared without violating its core privacy guarantee. Should viable alternative systematic disclosure avoidance methods be devised, it is important that they provide transparency without undoing important privacy protections. At the same time, we note that not all disclosure avoidance measures need be technical; documentation with some potential to increase disclosure risk could be protected with access and use restrictions, even if the dataset is public.

References

[1] Moustafa Alzantot and Mani Srivastava. 2019. Differential Privacy Synthetic Data Generation Using WGANs. https://github.com/nesl/nist_differential_privacy_synthetic_data_challenge.

[2] Raffael Bild, Klaus A. Kuhn, and Fabian Prasser. 2018. SafePub: A Truthful Data Anonymization Algorithm with Strong Privacy Guarantees. *PoPETs* 2018, 1 (2018). https://doi.org/10.1515/popets-2018-0004.

[3] Bill & Melinda Gates Foundation. 2024. Open Access Policy. https://openaccess.gatesfoundation.org/.

[4] danah boyd and Jayshree Sarathy. 2022. Differential Perspectives: Epistemic Disconnects Surrounding the U.S. Census Bureau's Use of Differential Privacy. *Harvard Data Science Review* Special Issue 2 (June 2022). https://hdsr.mitpress.mit.edu/pub/3vj5j6i0/release/3.

[5] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006). https://doi.org/10.1191/1478088706qp063oa.

[6] Mark Bun, Marco Carmosino, Palak Jain, Gabriel Kaptchuk, and Satchit Sivakumar. 2025. Enforcing Demographic Coherence: A Harms Aware Framework for Reasoning about Private Data Release. https://doi.org/10.48550/arXiv.2502.02709.

[7] Tânia Carvalho, Luís Antunes, Cristina Costa Santos, and Nuno Moniz. 2025. Empowering Open Data Sharing for Social Good: A Privacy-Aware Approach. *Scientific Data* 12 (2025). https://doi.org/10.1038/s41597-025-04506-x.

[8] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luís Antunes. 2023. Survey on Privacy-Preserving Techniques for Microdata Publication. *Comput. Surveys* 55, 14s (July 2023). https://doi.org/10.1145/3588765.

[9] Somiya Chhillar, Mary K. Righi, Rebecca E. Sutter, and Evgenios M. Kornaropoulos. 2025. Exposing Privacy Risks in Anonymizing Clinical Data: Combinatorial Refinement Attacks on k-Anonymity without Auxiliary Information. In *ACM CCS '25*. https://doi.org/10.1145/3719027.3765139.

[10] Aloni Cohen. 2022. Attacks on Deidentification's Defenses. In *USENIX Security '22*. https://www.usenix.org/conference/usenixsecurity22/presentation/cohen.

[11] DataCebo. 2026. SDMetrics. https://docs.sdv.dev/sdmetrics.

[12] DataCebo. 2026. Synthetic Data Vault. https://docs.sdv.dev/sdv.

[13] Cynthia Dwork. 2006. Differential Privacy. In *International Colloquium on Automata, Languages, and Programming (Lecture Notes in Computer Science)*. https://doi.org/10.1007/11787006_1.

[14] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. 2019. Differential Privacy in Practice: Expose Your Epsilons! *Journal of Privacy and Confidentiality* 9, 2 (Oct. 2019). https://doi.org/10.29012/jpc.689.

[15] Daniel Franzen, Saskia Nuñez von Voigt, Peter Sörries, Florian Tschorsch, and Claudia Müller-Birn. 2022. Am I Private and If So, How Many? Communicating Privacy Guarantees of Differential Privacy with Risk Communication Formats. In *ACM CCS '22*. https://doi.org/10.1145/3548606.3560693.

[16] Georgi Ganev, Meenatchi Sundaram Muthu Selva Annamalai, Sofiane Mahiou, and Emiliano De Cristofaro. 2025. The Importance of Being Discrete: Measuring the Impact of Discretization in End-to-End Differentially Private Synthetic Data. In *ACM CCS '25*. https://doi.org/10.1145/3719027.3765091.

[17] Georgi Ganev, Kai Xu, and Emiliano De Cristofaro. 2024. Graphical vs. Deep Generative Models: Measuring the Impact of Differentially Private Mechanisms and Budgets on Utility. In *ACM CCS '24*. https://doi.org/10.1145/3658644.3690215.

[18] Simson Garfinkel, Barbara Guttman, Joseph Near, Aref Dajani, and Phyllis Singer. 2023. *De-Identifying Government Datasets: Techniques and Governance*. NIST Special Publication 800-188. U.S. National Institute of Standards and Technology. https://csrc.nist.gov/pubs/sp/800/188/final.

[19] Elisabeth Gerber, Jeffrey Morenoff, and Mara Ostfeld. 2025. Detroit Metro Area Communities Study (DMACS) Wave 16, Michigan, 2023. https://doi.org/10.3886/ICPSR38892.v2.

[20] Wentao Guo, Aditya Kishore, Adam J Aviv, and Michelle L Mazurek. 2024. A Qualitative Analysis of Practical De-Identification Guides. In *ACM CCS '24*. https://doi.org/10.1145/3658644.3690270.

[21] Wentao Guo, Paige Pepitone, Adam J. Aviv, and Michelle L. Mazurek. 2025. How Researchers De-Identify Data in Practice. In *USENIX Security '25*. https://www.usenix.org/conference/usenixsecurity25/presentation/guo-wentao.

[22] Shlomi Hod and Ran Canetti. 2025. Differentially Private Release of Israel's National Registry of Live Births. In *IEEE S&P '25*. https://doi.org/10.1109/SP61157.2025.00101.

[23] Anco Hundepool, Ramya Ramaswamy, Peter-Paul de Wolf, Luisa Franconi, Ruth Brand, and Josep Domingo. 2021. μ-ARGUS. https://research.cbs.nl/casc/mu.htm.

[24] ICPSR. 2026. https://www.icpsr.umich.edu/sites/icpsr/home.

[25] Japan Science and Technology Agency. 2022. JST Policy on Open Access to Research Publications and Research Data Management. https://www.jst.go.jp/EN/about/openscience/policy_openscience_en_r4.pdf.

[26] Jiankai Jin, Eleanor McMurtry, Benjamin I. P. Rubinstein, and Olga Ohrimenko. 2022. Are We There Yet? Timing and Floating-Point Attacks on Differential Privacy Systems. In *IEEE S&P '22*. https://doi.org/10.1109/SP46214.2022.9833672.

[27] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. 2019. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In *ICLR '19*. https://openreview.net/forum?id=S1zk9iRqF7.

[28] Bogdan Kulynych, Theresa Stadler, Jean Louis Raisaro, and Carmela Troncoso. 2026. Should I Use Synthetic Data for That? An Analysis of the Suitability of Synthetic Data for Data Sharing and Augmentation. https://doi.org/10.48550/arXiv.2602.03791.

[29] Eric S. Lander. 2021. Public Access Congressional Report. https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf.

[30] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy beyond k-Anonymity and ℓ-Diversity. In *IEEE International Conference on Data Engineering*. https://doi.org/10.1109/ICDE.2007.367856.

[31] Qinyi Liu, Ronas Shakya, Jelena Jovanovic, Mohammad Khalil, and Javier de la Hoz-Ruiz. 2025. Ensuring Privacy through Synthetic Data Generation in Education. *British Journal of Educational Technology* 56 (2025). https://doi.org/10.1111/bjet.13576.

[32] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. ℓ-Diversity: Privacy beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (March 2007). https://doi.org/10.1145/1217299.1217302.

[33] Mike McIntire. 2026. Genetic Data from over 20,000 U.S. Children Misused for 'Race Science'. *The New York Times* (Jan. 2026). https://www.nytimes.com/2026/01/24/us/children-genetics-race-science.html.

[34] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. 2021. Winning the NIST Contest: A Scalable and General Approach to Differentially Private Synthetic Data. *Journal of Privacy and Confidentiality* 11, 3 (2021). https://doi.org/10.29012/jpc.778.

[35] Microsoft. 2026. Presidio. https://microsoft.github.io/presidio/.

[36] Kirsten N. Morehouse, Benedek Kurdi, and Brian A. Nosek. 2024. Responsible Data Sharing: Identifying and Remedying Possible Re-Identification of Human Participants. *American Psychologist* (2024). https://doi.org/10.1037/amp0001346.

[37] Jeffrey Morenoff and Mara Ostfeld. 2026. Detroit Metro Area Communities Study. https://detroitsurvey.umich.edu/.

[38] MOSTLY AI. 2026. Synthetic Data SDK. https://mostly-ai.github.io/mostlyai/.

[39] David Munechika, Zijie J. Wang, Jack Reidy, Josh Rubin, Krishna Gade, Krishnaram Kenthapadi, and Duen Horng Chau. 2022. Visual Auditor: Interactive Visualization for Detection and Summarization of Model Biases. In *VIS '22*. https://doi.org/10.1109/VIS54862.2022.00018.

[40] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. 2022. Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases. *PoPETs* 2022, 2 (2022). https://doi.org/10.2478/popets-2022-0058.

[41] Priyanka Nanayakkara, Elena Ghazi, and Salil Vadhan. 2026. Making Privacy Public: Toward a Differential Privacy Deployment Registry. In *IEEE S&P '26*. https://doi.org/10.1109/SP63933.2026.00157.

[42] Priyanka Nanayakkara and Jessica Hullman. 2023. What's Driving Conflicts around Differential Privacy for the U.S. Census. *IEEE Security & Privacy* 21, 5 (2023). https://doi.org/10.1109/MSEC.2022.3202793.

[43] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2023. What Are the Chances? Explaining the Epsilon Parameter in Differential Privacy. In *USENIX Security '23*. https://www.usenix.org/conference/usenixsecurity23/presentation/nanayakkara.

[44] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-Anonymization of Large Sparse Datasets. In *IEEE S&P '08*. https://doi.org/10.1109/SP.2008.33.

[45] Arvind Narayanan and Vitaly Shmatikov. 2010. Myths and Fallacies of "Personally Identifiable Information". *Commun. ACM* 53, 6 (June 2010). https://doi.org/10.1145/1743546.1743558.

[46] Alondra Nelson. 2022. Ensuring Free, Immediate, and Equitable Access to Federally Funded Research. https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.

[47] Ivoline C. Ngong, Brad Stenger, Joseph P. Near, and Yuanyuan Feng. 2024. Evaluating the Usability of Differential Privacy Tools with Data Practitioners. In *SOUPS '24*. https://www.usenix.org/conference/soups2024/presentation/ngong.

[48] NHS AI Lab Skunkworks. 2022. Synthetic Data Generation Pipeline. https://nhsx.github.io/skunkworks/synthetic-data-pipeline.

[49] Beata Nowok, Gillian M. Raab, and Chris Dibben. 2016. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software* 74, 11 (Oct. 2016). https://doi.org/10.18637/jss.v074.i11.

[50] Office of the Chief Science Advisor of Canada. 2020. Roadmap for Open Science. https://science.gc.ca/site/science/en/office-chief-science-advisor/open-science/roadmap-open-science.

[51] Paul Ohm. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review* 57 (2010). https://papers.ssrn.com/abstract=1450006.

[52] OpenAI. 2025. Whisper. https://openai.com/index/whisper/.

[53] OpenAIRE. 2025. Amnesia. https://amnesia.openaire.eu/.

[54] OpenDP. 2026. https://docs.opendp.org.

[55] Ileana Montoya Perez, Parisa Movahedi, Valtteri Nieminen, Antti Airola, and Tapio Pahikkala. 2024. Does Differentially Private Synthetic Data Lead to Synthetic Discoveries? *Methods of Information in Medicine* 63 (2024). https:

//doi.org/10.1055/a-2385-1355.
[56] Personal Data Protection Commission of Singapore. 2024. Data Anonymisation Tool. https://www.pdpc.gov.sg/organisations/resources/guidance-by-topic/basic-anonymisation.
[57] Fabian Prasser, Johanna Eicher, Helmut Spengler, Raffael Bild, and Klaus A. Kuhn. 2020. Flexible Data Anonymization Using ARX—Current Status and Challenges Ahead. *Software: Practice and Experience* 50, 7 (July 2020). https://doi.org/10.1002/spe.2812.
[58] pyannote. 2024. speaker-diarization. https://huggingface.co/pyannote/speaker-diarization.
[59] Zhaozhi Qian, Thomas Callender, Bogdan Cebere, Sam M. Janes, Neal Navani, and Mihaela van der Schaar. 2024. Synthetic Data for Privacy-Preserving Clinical Risk Prediction. *Scientific Reports* 14 (2024). https://doi.org/10.1038/s41598-024-72894-y.
[60] Gabriele Santangelo, Giovanna Nicora, Riccardo Bellazzi, and Arianna Dagliati. 2025. How Good Is Your Synthetic Data? SynthRO, a Dashboard to Evaluate and Benchmark Synthetic Tabular Data. *BMC Medical Informatics and Decision Making* 25 (2025). https://doi.org/10.1186/s12911-024-02731-9.
[61] Jayshree Sarathy and danah boyd. 2024. Statistical Imaginaries, State Legitimacy: Grappling with the Arrangements Underpinning Quantification in the U.S. Census. *Critical Sociology* (2024). https://doi.org/10.1177/08969205241270898.
[62] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. 2023. Don't Look at the Data! How Differential Privacy Reconfigures the Practices of Data Science. In *CHI '23*. https://doi.org/10.1145/3544548.3580791.
[63] Aniruddha Sen, Christine Task, Dhruv Kapur, Gary Howarth, and Karan Bhagat. 2023. Diverse Community Data for Benchmarking Data Privacy Algorithms. In *NeurIPS '23*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/a15032f8199511ced4d7a8e2bbb487a5-Abstract-Datasets_and_Benchmarks.html.
[64] Mary Anne Smart, Dhruv Sood, and Kristen Vaccaro. 2022. Understanding Risks of Privacy Theater with Differential Privacy. *PACMHCI* 6, CSCW2 (Nov. 2022). https://doi.org/10.1145/3555762.
[65] SmartNoise. 2026. Smartnoise-Synth. https://docs.smartnoise.org/.
[66] Mark F. St. John, Grit Denker, Peeter Laud, Karsten Martiny, Alisa Pankova, and Dusko Pavlovic. 2021. Decision Support for Sharing Data Using Differential Privacy. In *IEEE VizSec '21*. https://doi.org/10.1109/VizSec53666.2021.00008.
[67] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data – Anonymisation Groundhog Day. In *USENIX Security '22*. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.
[68] Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely. https://dataprivacylab.org/projects/identifiability/.
[69] Latanya Sweeney. 2002. *k*-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (Oct. 2002). https://doi.org/10.1142/S0218488502001648.
[70] Latanya Sweeney, Ji Su Yoo, Laura Perovich, Katherine E. Boronow, Phil Brown, and Julia Green Brody. 2017. Re-Identification Risks in HIPAA Safe Harbor Data: A Study of Data from One Environmental Health Study. *Technology Science* (2017). https://pmc.ncbi.nlm.nih.gov/articles/PMC6344041/.
[71] Christine Task, Karan Bhagat, and Gary Howarth. 2025. SDNist. https://doi.org/10.18434/mds2-2943.
[72] Matthias Templ. 2026. AI-assisted Statistical Disclosure Control with sdcMicro. https://cran.r-project.org/web/packages/sdcMicro/vignettes/ai_assisted_anonymization.html.
[73] Matthias Templ, Bernhard Meindl, Alexander Kowarik, Johannes Gussenbauer, Organisation For Economic Co-Operation And Development, Statistics Netherlands, and Pascal Heus. 2026. sdcMicro. https://cloud.r-project.org/web/packages/sdcMicro/index.html.
[74] The Counter Trafficking Data Collaborative. 2024. The Global Synthetic Dataset. https://www.ctdatacollaborative.org/page/global-synthetic-dataset.
[75] UK Research and Innovation. 2023. UKRI Open Access Policy. https://www.ukri.org/publications/ukri-open-access-policy/.
[76] United States Census Bureau. 2021. Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results. https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html.
[77] U.S. National Library of Medicine. 2019. NLM-Scrubber. https://lhncbc.nlm.nih.gov/scrubber/index.html.
[78] Kaitlyn R. Webb, Soumya Mukherjee, Aratrika Mustafi, Aleksandra Slavković, and Lars Vilhuber. 2026. Assessing Utility of Differential Privacy for RCTs. https://doi.org/10.48550/arXiv.2309.14581.
[79] Wellcome. 2024. Open Access Policy. https://wellcome.org/grant-funding/guidance/open-access-guidance/open-access-policy.
[80] World Health Organization. 2022. Sharing and Reuse of Health-Related Data for Research Purposes: WHO Policy and Implementation Guidance. https://www.who.int/publications/i/item/9789240044968.
[81] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Transactions on Database Systems* 42, 4 (Oct. 2017). https://doi.org/10.1145/3134428.
[82] Jonathan Zheng, Sauvik Das, Alan Ritter, and Wei Xu. 2025. Probabilistic Reasoning with LLMs for Privacy Risk Estimation. In *NeurIPS '25*. https://papers.nips.cc/paper_files/paper/2025/hash/9e0d77d3ce77fb7369663c27d7bcc838-Abstract-Conference.html.

Acknowledgments

This work was supported in part by NSF grant OAC-2232863. We are grateful to Eliana Kang for her help cleaning transcripts. We also thank Gabe Kaptchuk, Jessica Vitak, and Fumeng Yang for their valuable guidance and feedback. Paige Pepitone and Nathan Barrett volunteered their time to help us pilot this study—thank you!

Ethical considerations

As this study involved human participants and a restricted-use dataset, we took several steps to minimize risks. Our study was approved by the Institutional Review Boards of the University and Maryland and the George Washington University. Participants gave informed consent before taking part, including for session recordings and automated transcription. We only stored data locally and with third-party vendors whose contracts with our institutions include restrictions on use.

As noted in §3.2, the DMACS dataset used in the study is available to ICPSR members under terms of use. We complied with the applicable requirements and restrictions. We additionally obtained permission from the data authors to use this data for our study. All study participants were authorized users; they had access to the data only under our supervision and did not attempt re-identification.

To further reduce risks, we limit the reporting of details that could re-identify either study participants or individuals in the DMACS data. We do share some aggregate information about the data; in general, more detailed information is already publicly available on ICPSR in the dataset codebook.

Open science

The following materials are available at https://doi.org/10.5281/zenodo.22680856:

(1) DMACS dataset terms of use
(2) DMACS dataset codebook
(3) Pre-processing script used to prepare the DMACS dataset for sdcMicro and SmartNoise
(4) Survey questions associated with the variables in the modified DMACS dataset
(5) sdcMicro instruction handout
(6) SmartNoise/SDMetrics Jupyter notebook
(7) Qualitative codebook

In accordance with the terms of use, we cannot share the DMACS dataset used in this study. The data is available to authorized users through ICPSR (study 38892): https://www.icpsr.umich.edu/web/ICPSR/studies/38892.

As alluded to in §3.1, we slightly modified SDMetrics to fix some distracting usability issues. Specifically, we made visualizations display ordered categories in the correct order, hid an irrelevant visualization, and made display of labels more consistent. The modified code is available at https://github.com/oatnewguo/SDMetrics.

AI use

No AI tools were used to write this paper. We used ChatGPT to generate illustrative free-text responses to replace redacted values in the DMACS dataset for one variable; this did not play a major role in our study, and the specific generated text can be found in the pre-processing script linked in the Open Science appendix. As noted in §3.5, we used AI tools (Whisper [52] and pyannote [58]) to create initial transcripts of recorded user studies; we listened through each recording to fix transcription mistakes.

A sdcMicro study protocol

*Mirror the computer screen with a larger display in the room so you can observe the participant's actions without looking over their shoulder. Open sdcApp. Load the data, and under the **Anonymize** tab mark the following variables as categorical key variables: **birth_year**, and every variable from **born_in_US** to the end. Separately, load and view the dataset in RStudio.*

Orientation

I'll start with some background. Publishing research data has benefits, including enabling replication and meta-analysis, and it is sometimes required by funding agencies and journals. To protect individuals' privacy, published data is often de-identified: that is, modified to make it harder to learn information that could be linked back to specific individuals in the real world. Some steps are straightforward—like removing names—but indirect identifiers can be complicated. For example, if a published dataset describes a female dentist born in 1993 who lives in College Park, then someone might be able to re-identify her and learn new information that should have been private.

To improve privacy, we can remove or alter data that could be identifying or sensitive. However, this could also limit the dataset's usefulness for legitimate purposes, creating trade-offs between privacy and utility. Any questions so far?

- Have you done anything like this (de-identifying and/or publishing research data) before?

Today we'll be working with a dataset modified from a 2023 survey of the Detroit Metro Area. This survey measured 2,000 residents' opinions and experiences relating to health, economics, and more. The overall goal is to inform local community development.

Please take a minute to familiarize yourself with the dataset in front of you. If you want to know more about the survey questions or the responses, I have more information. As you look through the data, please think out loud about what you think the use cases and privacy risks might be.

Participant looks through the dataset. If they ask, you can provide the questions and/or codebook. Take note of the use cases they mention, because we will ask them about them later.

- Suppose you've collected this data and are required to release it to the public. If you had access to an expert in de-identification, what would you want help with? What questions would you like to ask them?

Tool use

Today you will use a tool called sdcMicro to help de-identify this dataset. It's important for you to know we have no affiliation with sdcMicro: this is a tool developed by other researchers, and one of our main goals is to identify issues with it from the perspective of researchers like you, so we'd love to have your honest opinions, positive or negative.

sdcMicro tries to prevent individuals in the data from being re-identified—recall the example from before of the female dentist born in 1993 who lives in College Park. To do this, sdcMicro will apply an algorithm that identifies cases like this and then deletes individual values to reduce re-identifiability: for example, removing her city or occupation.

Give instructions to participant.

sdcMicro uses a risk metric called k-anonymity to decide how to de-identify data. To calculate k-anonymity, we first choose a set of variables that could be used to re-identify individuals. For example, we might pick age, gender, and city because they can be learned from voter registration records. Then, based on these variables, we determine whether individuals look different from others in the dataset. In the mini example on the left, the bottom four records are each unique; after deleting some values, on the right, there are two groups of similar-looking records, one with four men and one with three women. The value k is the size of the smallest of these groups, and higher values imply stronger privacy. Any questions?

Now, please follow the instructions to apply the de-identification algorithm and review the results. As you go along, please think out loud—whether it's something you find helpful or limiting, a connection to your own research experiences, or any other reaction. Go ahead whenever you're ready!

Ask follow-up questions and remind participants to think aloud as necessary. When they reach step 6, ask the questions below.

- What is your impression of the utility and privacy of this dataset in its current state? It's okay if you're not sure yet—we're also interested in what you're uncertain about.
  - In your own words, could you please try to explain to me how the data has been transformed?
    - Could you please describe for me how you expect the distribution of values has changed across the dataset, for single variables and crosstabs?
    - Could you please describe for me how you expect individual records to have changed due to de-identification?
  - As an example of a use case for this data, suppose a group of researchers is compiling datasets from multiple cities to analyze mental health trends across demographics. How do you think the current de-identification has impacted the usefulness of this data for that study?
    - *Alternatively, ask them about use cases they described at the beginning of the study.*

    - As an example of a privacy threat to this data, suppose a local health insurance provider knows people's demographics and wants to use this data to learn about individuals' disabilities and finances. How do you think the current de-identification has impacted the protection of this data against that threat?
      - *Alternatively, ask them about privacy risks they described at the beginning of the study.*

[if time permits] We have some extra time, so let's try de-identifying the dataset again. Step 7 of your instructions lists some suggestions for something to do differently. I can help you with this. Is there any option that you'd like to try? *[Optionally, suggest an option based on participant's earlier comments.]*

*Reset the anonymization using **Reset SDC problem** under the **Anonymize** tab. Guide the participant through their choice.*

- What is your impression of how the utility and privacy of this dataset are different than before?
  - In the context of publishing de-identified research data, which version of the dataset do you prefer?

Reflection

Let's talk about your overall experience and impressions. Please remember that we are not affiliated with sdcMicro at all, so we're really looking for your honest opinions, positive and negative.

- What is your overall impression of sdcMicro? Did it provide you with any information or capabilities that you would want in a tool for making research data less identifiable?
  - Compared to how you would like to be informed about the impact of de-identification, how was sdcMicro similar or different? Did you find it effective?
- To what extent do you believe that systematically deleting individual values, like sdcMicro did, creates difficulties for downstream use cases?
  - How do you believe these changes would affect attempts to replicate or conduct new analyses of this data?
  - How do you believe these changes would affect attempts to include this data in large-scale meta-analyses?
- As a researcher, can you think of situations where a tool like sdcMicro would be a good fit, or—on the other hand—situations where using sdcMicro would be a bad idea?
- When analyzing datasets with deleted values, what additional steps or considerations would someone need to take in order to account for these changes, in your opinion?
- If you were preparing to publish the kind of data you typically work with using a tool like sdcMicro, do you think there would be different needs or considerations than the ones we discussed today?
- Before we wrap up, is there anything else you want to say about the design of tools and methods for making research data less identifiable?

Thank you so much for your time and insights! I'll turn off the recording. Do you have any questions for me about this research?

B SmartNoise/SDMetrics study protocol

*Mirror the computer screen with a larger display in the room so you can observe the participant's actions without looking over their shoulder. From inside the virtual environment, run **jupyter lab notebook.ipynb** and minimize the file browser at the left (if applicable). Separately, load and view the dataset in RStudio.*

Orientation

I'll start with some background. Publishing research data has benefits, including enabling replication and meta-analysis, and it is sometimes required by funding agencies and journals. To protect individuals' privacy, published data is often de-identified: that is, modified to make it harder to learn information that could be linked back to specific individuals in the real world. Some steps are straightforward—like removing names—but indirect identifiers can be complicated. For example, if a published dataset describes a female dentist born in 1993 who lives in College Park, then someone might be able to re-identify her and learn new information that should have been private.

To improve privacy, we can remove or alter data that could be identifying or sensitive. However, this could also limit the dataset's usefulness for legitimate purposes, creating trade-offs between privacy and utility. Any questions so far?

- Have you done anything like this (de-identifying and/or publishing research data) before?

Today we'll be working with a dataset modified from a 2023 survey of the Detroit Metro Area. This survey measured 2,000 residents' opinions and experiences relating to health, economics, and more. The overall goal is to inform local community development.

Please take a minute to familiarize yourself with the dataset in front of you. If you want to know more about the survey questions or the responses, I have more information. As you look through the data, please think out loud about what you think the use cases and privacy risks might be.

Participant looks through the dataset. If they ask, you can provide the questions and/or codebook. Take note of the use cases they mention, because we will ask them about them later.

- Suppose you've collected this data and are required to release it to the public. If you had access to an expert in de-identification, what would you want help with? What questions would you like to ask them?

Tool use

Today you will use a tool called SmartNoise to generate a synthetic version of this dataset. It's important for you to know we have no affiliation with SmartNoise: this is a tool developed by other researchers, and one of our main goals is to identify issues with it from the perspective of researchers like you, so we'd love to have your honest opinions, positive or negative.

Direct participant to Jupyter Notebook.

You will be working from a document called a notebook. Please read and complete the part marked **Introduction**.

Wait for participant.

To protect individuals' privacy, SmartNoise will attempt to generate fake data that hides information about individuals while preserving characteristics of the dataset as a whole. This synthetic data would be released instead of the real data. Please follow the step labeled **Generating synthetic data** in your notebook to start the program; while it runs, I will explain more.

Generating synthetic data without a rigorous privacy framework could expose sensitive details about the original data. We want to preserve overall correlations—for example, relationships between race and gender and health outcomes—but if our correlations are too specific, we might reveal private information about individuals. An adversary who knows the computer program that was used and how it works might be able to reverse engineer private information even more effectively. Therefore, SmartNoise uses differential privacy, which is a framework that mathematically guarantees a certain level of privacy protection for each individual in the original data—even against an adversary who knows how SmartNoise works, has perfect information about the rest of the original dataset, and uses optimal techniques for guessing private information. It does this by automatically detecting correlations in the original data, adding some statistical noise to ensure those correlations aren't too specific, and then sampling synthetic data based on those fuzzy correlations.

The degree of privacy protection can be customized using a value called epsilon. Lower values of epsilon provide stronger privacy but distort the synthetic data more. Higher values produce more realistic synthetic data, but they enable adversaries to make better educated guesses. Epsilon can be any positive number, with real-world values ranging from 0.1 all the way to 50, though choices at either end mean severe trade-offs to accuracy or privacy. Choosing an appropriate epsilon depends on context. Real-world synthetic datasets—such as nationwide data about new infancies and data about demographics and experiences of human trafficking victims—have used an epsilon of about 10. Any questions?

To start, you have just generated synthetic data with an epsilon of 5.

Direct their attention to where epsilon is specified in the notebook.

We may have time to revisit this choice later. Now, please follow the instructions in your notebook to review the results. You do not need to read or understand the code, though some instructions will prompt you to modify some values; these values will always be at the top of their code block, and I can help if anything is confusing. As you go along, please think out loud—whether it's something you find helpful or limiting, a connection to your own research experiences, or any other reaction. Go ahead whenever you're ready!

Ask follow-up questions and remind participants to think aloud as necessary. When they reach step 6, ask the questions below.

- What is your impression of the utility and privacy of this dataset in its current state? It's okay if you're not sure yet—we're also interested in what you're uncertain about.
  - In your own words, could you please try to explain to me how the synthetic data differs from the original data?
    - Could you please describe for me how you expect the distribution of values has changed across the dataset, for single variables and crosstabs?
    - Could you please describe for me how you expect individual records in the synthetic data were created?
  - As an example of a use case for this data, suppose a group of researchers is compiling datasets from multiple cities to analyze mental health trends across demographics. How do you think generating this synthetic dataset has impacted the usefulness of this data for that study?
    - *Alternatively, ask them about use cases they described at the beginning of the study.*
  - As an example of a privacy threat to this data, suppose a local health insurance provider knows people's demographics and wants to use this data to learn about individuals' disabilities and finances. How do you think generating this synthetic dataset has impacted the protection of this data against that threat?
    - *Alternatively, ask them about privacy risks they described at the beginning of the study.*

[if time permits] We have some extra time, so let's try generating synthetic data again. Go back to the start and change the value of epsilon. Then, go through the notebook again. *[If they were okay with the data, make sure they choose a meaningfully lower epsilon.]*

- What is your impression of how the utility and privacy of this dataset are different than before?
  - In the context of publishing de-identified research data, which version of the dataset do you prefer?

Reflection

Let's talk about your overall experience and impressions. Please remember that we are not affiliated with SmartNoise at all, so we're really looking for your honest opinions, positive and negative.

- What is your overall impression of SmartNoise? Did it provide you with any information or capabilities that you would want in a tool for making research data less identifiable?
  - Compared to how you would like to be informed about the impact of generating synthetic data, how was SmartNoise similar or different? Did you find it effective?
- To what extent do you believe that generating synthetic data, like SmartNoise did, creates difficulties for downstream use cases?
  - How do you believe these changes would affect attempts to replicate or conduct new analyses of this data?
  - How do you believe these changes would affect attempts to include this data in large-scale meta-analyses?
- As a researcher, can you think of situations where a tool like SmartNoise would be a good fit, or—on the other hand—situations where using SmartNoise would be a bad idea?
- When analyzing synthetic datasets, what additional steps or considerations would someone need to take in order to account for changes from the real data, in your opinion?
- If you were preparing to publish the kind of data you typically work with using a tool like SmartNoise, do you think

there would be different needs or considerations than the ones we discussed today?
- Before we wrap up, is there anything else you want to say about the design of tools and methods for making research data less identifiable?

Thank you so much for your time and insights! I'll turn off the recording. There's one thing I need to clarify. I said the tool you used is called SmartNoise, which is not totally true. Actually, SmartNoise generated the synthetic data, and a different tool called SDMetrics created the visualizations and reports that you saw. Now, do you have any questions for me about this research?

C Screening questionnaire

Institutional affiliation

(1) Which institution are you primarily affiliated with?
- University of Maryland, College Park | The George Washington University | Other _____

(2) Which department are you primarily affiliated with? _____

(3) What is your role at your institution?
- Master's student | Doctoral student | Postdoctoral fellow | Faculty member | Other _____

(4) What is your specific research focus, if you have one? _____

Research experience

We are interested in learning more about your experience working with **quantitative human-subjects microdata**.

By **quantitative**, we mean structured data collected from hundreds or more research participants—for example, through surveys. By **human-subjects**, we mean data about people. And by **microdata**, we mean data that is disaggregated at the individual level (for example, each line in the dataset represents one person), as opposed to aggregate statistics.

Please answer the following questions to help us understand your background. You may provide additional context in the text boxes if you wish.

(1) How many graduate-level classes have you taken in which you analyzed **quantitative human-subjects microdata**? _____

(2) How many times have you analyzed **quantitative human-subjects microdata** for a research paper or report that was submitted to a peer-reviewed publication or research funder? _____

(3) How many times have you de-identified **quantitative human-subjects microdata** to prepare it for publication (e.g., in an online data archive)? _____

(4) How, if at all, have you learned about de-identifying **quantitative human-subjects microdata**? (Check all that apply.)
- In a class or professional training
- By reading guides, papers, or other literature
- By learning from collaborators and mentors or by doing it myself
- I have not learned about de-identifying quantitative human-subjects microdata

(5) Do you have other experience working with **quantitative human-subjects microdata** or de-identification that you'd like to share? _____

Participant demographics

(1) What is your age?
- 18–24 | 25–34 | 35–44 | 45–54 | 55–64 | 65+ | Prefer not to state

(2) What is your gender?
- Male | Female | Non-binary | Prefer not to state | Prefer to self-describe _____

(3) What is your race? (You can select more than one option.)
- American Indian or Alaska Native | Asian | Black or African American | Hispanic or Latino | Native Hawaiian or Pacific Islander | White | Prefer not to state | Prefer to self-describe _____

(4) How did you learn about this study?
- Message from an email listserv, research lab, or other group
- In-class announcement
- Word of mouth from another participant in this study
- Direct email from the researchers
- Other _____

(5) What is your email? We will use this in order to contact you about study logistics. _____

D Participant information

Table 3: These tables describe each participant's research area, broken down by which tool they used in our study.

(a) sdcMicro

	Research area
M1	Crime / justice
M2	Crime / justice
M3	Crime / justice
M4	Psychology
M5	Psychology
M6	Psychology
M7	Health / medicine
M8	Health / medicine
M9	Health / medicine
M10	Education
M11	Sociology
M12	Political science

(b) SmartNoise

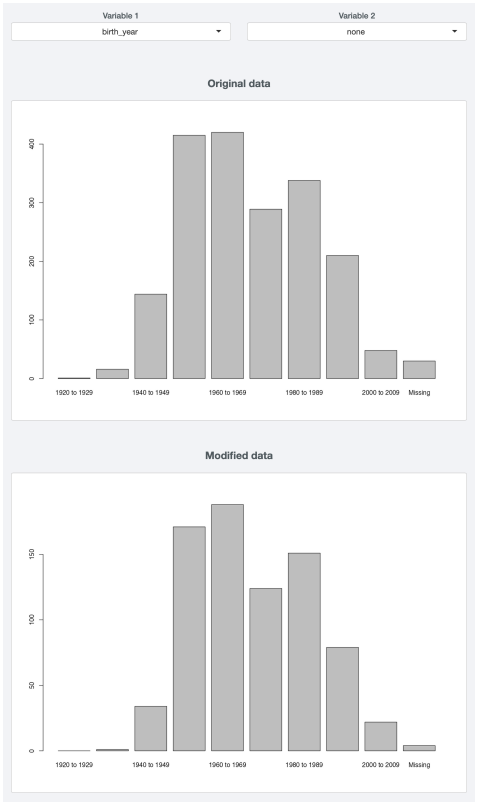
	Research area
N1	Crime / justice
N2	Crime / justice
N3	Crime / justice
N4	Psychology
N5	Psychology
N6	Psychology
N7	Health / medicine
N8	Health / medicine
N9	Health / medicine
N10	Health / medicine
N11	Education

Table 4: This table summarizes participants' demographics, as self-reported in the screening questionnaire.

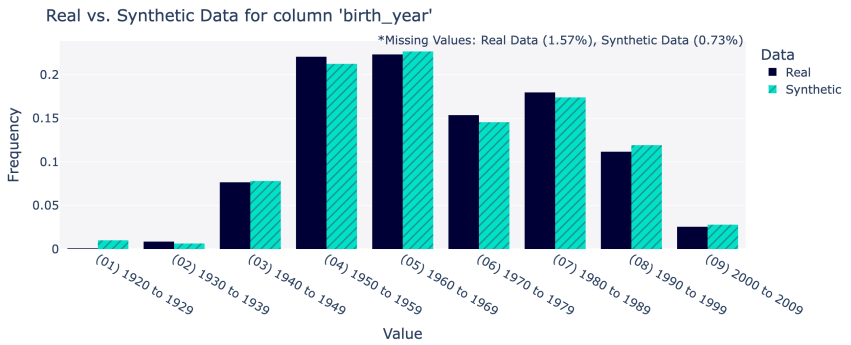
Category	Count
Gender	
Female	18
Male	5
Race	
White	12
Asian	7
Black or African American	3
Hispanic or Latino	1
Age	
18–24	3
25–34	16
35–44	4

E Examples of tool interfaces

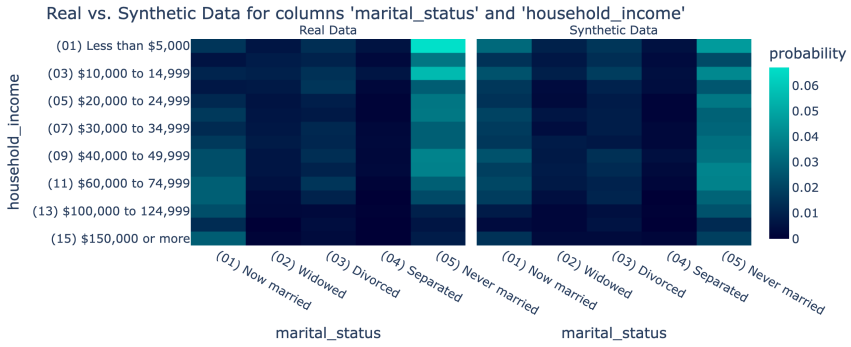
Figure 2: These screenshots are illustrative examples of tool interfaces that participants used to evaluate the utility and privacy of de-identified data. Complete study protocols, detailing all interfaces participants interacted with, are provided in Appendices A and B.



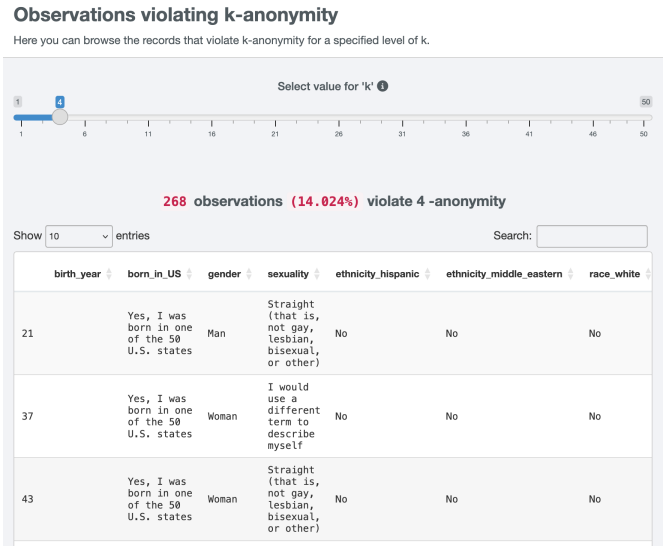
(a) sdcMicro: changes in univariate distributions



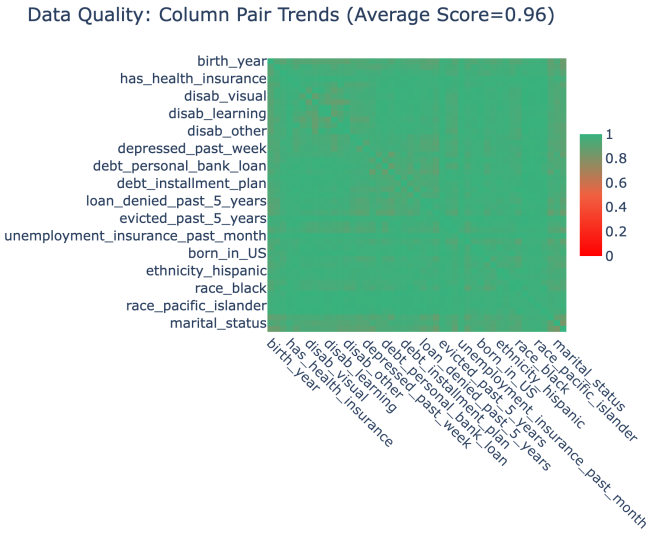
(b) SDMetrics: changes in univariate distributions



(c) SDMetrics: changes in crosstab distributions



(d) sdcMicro: records violating k-anonymity for higher values of k



(e) SDMetrics: visualization of all crosstab similarity scores