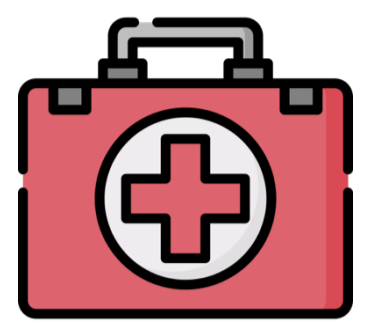


Helping Researchers De-Identify Data for Open Science

Wentao Guo, Wellington Barbosa,* Paige Pepitone,† Adam Aviv,* Michelle Mazurek
 University of Maryland, *The George Washington University, †NORC at the University of Chicago

✉ wguo5@umd.edu
 🦋 @wentaoguo.bsky.social
 🌐 wentaoguo.com

Motivating examples



Medical researchers publish **clinical trial** data.

Scientists verify the **safety** of new treatments.

But data on **physical and mental health** could leak to insurance companies.



Aid organizations publish data about **program outcomes**.

Journalists cover the **impact** of taxpayer-funded programs.

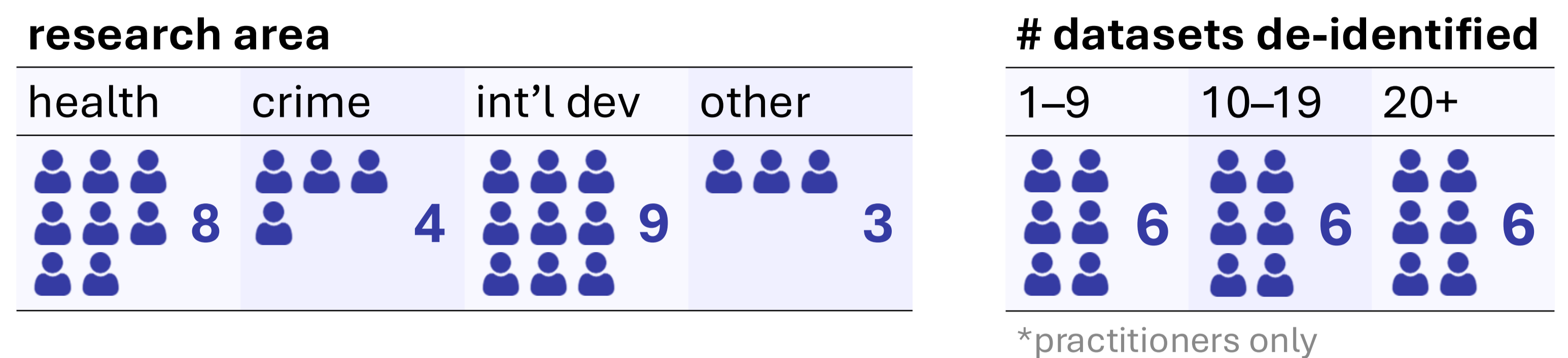
But data on **political sentiments** could leak to local organized crime groups.

Social, medical, & behavioral scientists are increasingly **required** to de-identify and publish data, despite the **difficulty of managing re-identification risk**.

Semi-structured interviews

We interviewed...

- **18 practitioners** who had de-identified and published research data
- **6 curators** who review data submissions for repositories



- RQ1.** How do researchers perceive **re-identification threats**?
- RQ2.** How do they **de-identify** data in practice?
- RQ3.** What **challenges** do they encounter?

icons: flaticon.com

RQ1 and RQ2. Mismatch between risk model and actual de-identification

Researchers are concerned about **combinations of indirect identifiers** that could link individuals to external data.

In practice, researchers search for **distinctive values** and combinations of values. However, most only inspect **pairwise combinations of identifiers** (at most) and rely on **informal and social processes** for evaluating success.

“You want to avoid putting clinicians into a **group of less than five similar clinicians**. Like, a 35-year-old Black endocrinologist from [a specific town]—there’s probably just one.”

“There might be a census block that **links back** to an external dataset. They’re **one of now like 200** people.”

quotes edited for brevity

1. Suppose we decide **age × occupation** is a particularly identifying combination.

2. Calculate crosstabs (2-way counts):

	18-24	25-29	30-34	...
Dentist	1	6	17	
Surgeon	0	2	7	

3. Some counts are too low! Let’s combine all three age categories into 18-34.

4. Repeat with different identifiers.

No evaluation of uniqueness by **age × occupation × race × gender × income × ...**

“I get a bit into the weeds sometimes, and I’m like, “Ooh, they have two chickens, and **nobody else has two chickens**.” And my boss is like, “Don’t worry about it; there’s a **very minute possibility** that somebody would go to this village, and they probably have more chickens now.”

“You could crosstab all variables in theory, but that would be like millions of crosstabs. Maybe it’s somebody’s **position, crosstabbed with their age** or gender. It’s not necessarily a scientific process. It’s **more knowing what to look for**.”

Why the mismatch?

1. Threats are seen as **unrealistic**.
2. Subsamples both mitigate risk and complicate de-ID.
3. Utility trade-offs are **unacceptable**.
4. Support and incentives are insufficient.

“I think **it is possible in many clinical datasets** to identify an individual, but the level of sophistication and effort you would need is **beyond the real threat**.”

“We **really struggle with dates and time**. Every time you apply a date shift, you **severely limit the value** of your data.”

Curator:
 “Data submitters can propose an access level, but it doesn’t really matter, because **the repository has the final say**.”



Practitioner:
 “The data was basically **rendered useless** by the amount of de-identification we had to do. I could say I want the highest level of security, but **they don’t have to do what I say**.”

More of our work

How researchers de-identify data in practice. *USENIX '25*.

<https://www.usenix.org/conference/usenixsecurity25/presentation/guo-wentao>

A qualitative analysis of practical de-identification guides. *CCS '24*.

<https://www.usenix.org/conference/usenixsecurity25/presentation/guo-wentao>

An exploratory user study of disclosure avoidance tools for scientific microdata. *In submission*.

- User study with 23 scientists de-identifying a real research dataset, using one of two disclosure avoidance tools:
 - sdcMicro: (*k*-anonymity)
 - SmartNoise + SDMetrics: (differential privacy)
- Skepticism about baseline re-identification risk causes systematic disclosure avoidance methods to be seen as **extreme**
- High-level metrics + low-level visual comparisons = scientists can **iteratively build a stronger understanding** of data utility
- Tools are **not very effective** at communicating info about privacy

