

# Your Data Are a Fingerprint

*Why Anonymization is Not Anonymous (and how Statistics can protect you)*

Dylan Spicker

January 31, 2025

A tremendous amount of data  
are collected by researchers,  
companies, governments, and  
similar institutions.

# Pillar #1

Learning from data is  
crucial for our  
understanding of the  
world.

## Pillar #1

Learning from data is crucial for our understanding of the world.

## Pillar #2

Protecting individual privacy is necessary as data become more abundant.

Imagine *someone* trying to  
learn your private information  
using information based on  
your data.

This is analogous to a detective  
searching a crime scene.

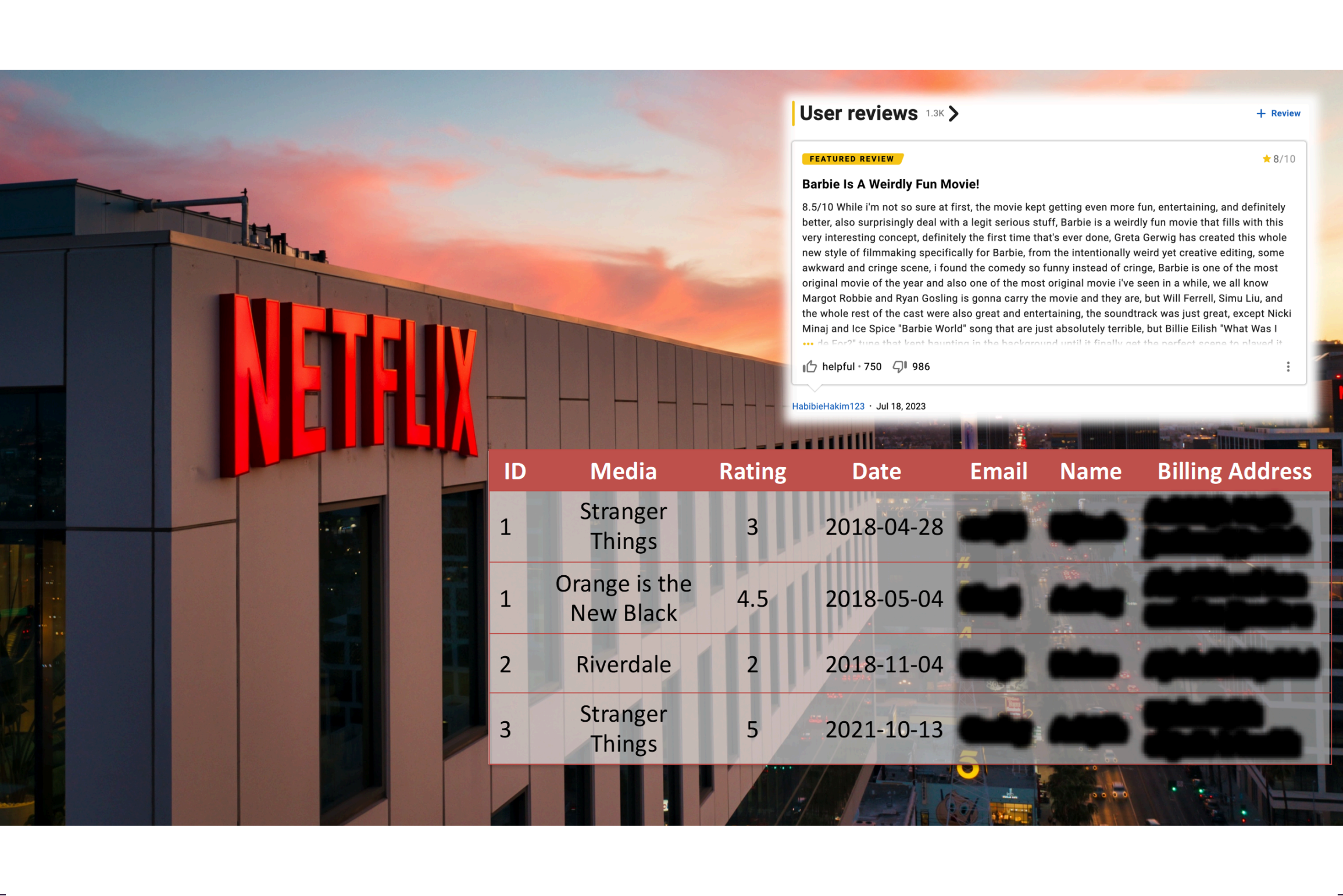
id	first_name	last_name	gender	prescription
03-2310	Fancy	Thonason	Female	Diclofenac Sodium
01-7096	Tedd	Hanton	Male	Octinoxate and Oxybenzone
05-6875	Ashlin	Byatt	Male	Permethrin
08-0157	Ashby	Pughe	Male	Nafcillin
04-2748	Avictor	Salerg	Male	Naproxen Sodium
09-3263	Court	O'Deegan	Male	OCTINOXATE, OXYBENZONE, ZINC OXIDE
09-4225	Wainwright	Marlor	Male	DIMETHICONE
07-6259	Renelle	Byk	Female	Clobetasol Propionate
07-6699	Greggory	Dannett	Female	standardized senna concentrate and docusate sodium
00-8977	Hedvige	Barkus	Female	Diphenhydramine Hydrochloride
02-2428	Martie	Kinzel	Male	leum
09-1507	Fairfax	Wintersgill	Male	lamotrigine
01-7104	Dulce	Tomankiewicz	Female	risperidone
04-6274	Adamo	Mc Caughan	Male	OPERCULATA FRUIT and MERCURIC IODIDE and SUS SCROFA NASAL ML
00-1324	Odey	Santus	Male	Levofloxacin
01-9531	Beck	Bullier	Genderfluid	Menthol
02-5248	Nita	Castagnone	Female	Acetaminophen
06-5308	Betta	Vedstra	Female	Oxycodone Hydrochloride and Aspirin
00-9924	Sharona	Pallaske	Female	Salicylic acid
04-7089	Wash	M'Quhan	Male	levalbuterol hydrochloride

## Caught “Red-Handed”: Direct Data Releases

If data are **directly released** this constitutes a **complete privacy violation**.

What if data are **anonymized**?

Data anonymization is the process of removing clear personal identifiers from a specific dataset.



User reviews 1.3K >

+ Review

FEATURED REVIEW

★ 8/10

Barbie Is A Weirdly Fun Movie!

8.5/10 While I'm not so sure at first, the movie kept getting even more fun, entertaining, and definitely better, also surprisingly deal with a legit serious stuff, Barbie is a weirdly fun movie that fills with this very interesting concept, definitely the first time that's ever done, Greta Gerwig has created this whole new style of filmmaking specifically for Barbie, from the intentionally weird yet creative editing, some awkward and cringe scene, i found the comedy so funny instead of cringe, Barbie is one of the most original movie of the year and also one of the most original movie i've seen in a while, we all know Margot Robbie and Ryan Gosling is gonna carry the movie and they are, but Will Ferrell, Simu Liu, and the whole rest of the cast were also great and entertaining, the soundtrack was just great, except Nicki Minaj and Ice Spice "Barbie World" song that are just absolutely terrible, but Billie Eilish "What Was I ... do For?" tune that kept haunting in the background until it finally got the perfect scene to played it

👍 helpful · 750    🗨️ 986

⋮

HabibieHakim123 · Jul 18, 2023

ID	Media	Rating	Date	Email	Name	Billing Address
1	Stranger Things	3	2018-04-28	[REDACTED]	[REDACTED]	[REDACTED]
1	Orange is the New Black	4.5	2018-05-04	[REDACTED]	[REDACTED]	[REDACTED]
2	Riverdale	2	2018-11-04	[REDACTED]	[REDACTED]	[REDACTED]
3	Stranger Things	5	2021-10-13	[REDACTED]	[REDACTED]	[REDACTED]



Sufficient amounts of non-  
personally identifying  
information can act as personal  
identifiers.

Summary statistics like means, counts, or standard deviations can be reported to describe the overall data.



Filters



5947  
Results

0 Filters

Search for a filter or table

Geographies

Nation

State

County

County Subdivision

Place

ZIP

Metropolitan Area

State

County

Block

Block

All Geographies

Topics

Business and Economics

Education

Employment

Families and Living Arrangements

Government

Health

Housing

Income and Poverty

Populations and People

Race and Ethnicity

Surveys

American Community Survey

Community Resilience Estimates

Current Population Survey

Decennial Census

Decennial Census of Island Areas

Economic Census

Economic Census of Island Areas

Economic Surveys

Geography

Household Pulse Survey

International Database

Population Estimates

Post-Secondary Employment

5947 Results

View: 10 | 25 | 50

Decennial Census

P1 | TOTAL POPULATION

View All 18 Products

American Community Survey

S0101 | Age and Sex

View All 27 Products

2023: ACS 1-Year Estimates Subject Tables

S0101 | AGE AND SEX

American Community Survey | 2018: ACS 5-Year Estimates Subject Tables

Label	United States		Percent		Male		Percent Male	
	Total							
	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error
▼ Total population	322,903,030	*****	(X)	(X)	158,984,190	±6,691	(X)	(X)
▼ AGE								
Under 5 years	19,836,850	±4,121	6.1%	±0.1	10,146,960	±3,495	6.4%	±0.1
5 to 9 years	20,311,494	±21,979	6.3%	±0.1	10,357,512	±14,342	6.5%	±0.1

# Data “Finger-Printing”: Summary Statistic Releases

Given **summary statistics**, the US Census Bureau demonstrated that **reconstruction attacks** are possible (and **astonishingly accurate**).

10 to 14 years	20,817,510	±5,580	6.5%	±0.1	10,357,512	±13,725	6.5%	±0.1
15 to 19 years	20,261,694	±5,788	6.3%	±0.1	10,899,332	±13,985	6.7%	±0.1
20 to 24 years	21,811,374	±21,767	6.8%	±0.1	10,405,721	±15,714	6.8%	±0.1
25 to 34 years	25,124,444	±24,235	7.8%	±0.1	12,472,785	±14,225	9.8%	±0.1
35 to 44 years	25,549,216	±15,459	7.9%	±0.1	13,813,532	±18,706	9.4%	±0.1
45 to 54 years	24,928,433	±14,458	7.7%	±0.1	13,421,232	±18,706	9.4%	±0.1
55 to 64 years	21,028,815	±6,071	6.5%	±0.1	11,001,866	±4,461	13.3%	±0.1
65 to 74 years	19,367,812	±4,121	6.0%	±0.1	10,476,243	±3,286	10.6%	±0.1
75 years and over	12,852,362	±5,427	4.0%	±0.1	6,544,568	±5,784	10.9%	±0.1
18 years and over	257,754,672	±18,355	79.8%	±0.1	125,899,327	±11,125	78.1%	±0.1
60 years and over	236,122,501	±26,806	73.1%	±0.1	114,611,937	±15,504	72.1%	±0.1
60 years and over	68,913,938	±20,510	21.3%	±0.1	31,209,043	±15,769	19.6%	±0.1
62 years and over	60,628,688	±17,851	18.8%	±0.1	27,221,257	±13,204	17.1%	±0.1
65 years and over	49,238,581	±5,463	15.2%	±0.1	21,781,300	±3,215	13.7%	±0.1
75 years and over	20,703,162	±3,426	6.4%	±0.1	8,442,251	±2,519	5.3%	±0.1
▼ SUMMARY INDICATORS								
Median age (years)	37.9	±0.1	(X)	(X)	36.6	±0.1	(X)	(X)
Sex ratio (males per 100 females)	97.0	±0.1	(X)	(X)	(X)	(X)	(X)	(X)
Age dependency ratio	61.4	±0.1	(X)	(X)	(X)	(X)	(X)	(X)
Old-age dependency ratio	24.6	±0.1	(X)	(X)	(X)	(X)	(X)	(X)
Child dependency ratio	36.8	±0.1	(X)	(X)	(X)	(X)	(X)	(X)
▼ PERCENT ALLOCATED								
Sex	(X)	(X)	0.1%	(X)	(X)	(X)	(X)	(X)

You leave **trace amounts** of  
personal information in  
**summary statistics.**

What about using **advanced**  
statistical procedures?

ORIGINAL ARTICLE

Differentially private outcome-weighted least squares dynamic treatment regime estimation

Dylan Spicker<sup>1</sup> | Erica E. M. Moodie<sup>2</sup> | Susan M. Shortt<sup>3</sup>

<sup>1</sup>Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, New Brunswick, Canada

<sup>2</sup>Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada

<sup>3</sup>Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA

<sup>4</sup>Department of Biostatistics, University of Washington, Seattle, WA, USA

Correspondence

Dylan Spicker, Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, NB, Canada.  
Email: dylan.spicker@unb.ca

Funding Information

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01 MH114873. This research was undertaken, in part, thanks to funding from the Canada Excellence Research Chairs Program. Additionally, Dylan Spicker was supported by the Canadian Statistical Sciences Institute (CANSI) and STATLAB and Erica E.M. Moodie by a chercheur de mérite career award from the FRQ, Santé. The STAR\*D study was supported by NIMH Contract #N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is CT00021528.

Precision medicine is a framework for developing evidence-based medical recommendations that seeks to determine the optimal sequence of treatments, tailored to all of the relevant, observable patient-level characteristics. Because precision medicine relies on highly sensitive, patient-level data, the privacy of this information is of great importance. Dynamic treatment regime estimation (DTR) is a family of techniques for estimating optimal treatment policies in a longitudinal setting. Outcome-weighted least squares (OWLS) techniques leverage support vector machines (SVMs) to perform estimation. SVMs perform classification based on a set of influential points in the data known as support vectors. The classification rule produced by SVMs often requires direct access to the support vectors. This is a concern because releasing a treatment policy estimated with OWL requires the release of patient data for a subset of patients in the sample. As a result, the classification rules from SVMs constitute a severe privacy violation for those individuals whose data comprise the support vectors. This privacy violation is a major concern, particularly in light of the potentially highly sensitive medical data that are used in DTR estimation. Differential privacy has emerged as a mathematical framework for ensuring the privacy of individual-level data, with provable guarantees on the likelihood that individual data can be determined by an adversary. We provide the first investigation of DTRs and provide a differentially private OWLS estimator, with theoretical results allowing us to quantify the cost of privacy in terms of the accuracy of the private estimators.

KEYWORDS

differential privacy, dynamic treatment regimes, individualized treatment rules, precision medicine, support vector machines

# Circumstantial Evidence: Advanced Statistical Modelling

Even advanced statistical modelling is susceptible to data leakage. The severity depends on the underlying analysis.

## 1 | INTRODUCTION

Health data are, by their nature, sensitive information. We assume that patients expect that their sensitive information will remain protected and secure once it has been collected. The expectation is that no one outside of those who were explicitly granted access (such as the researchers receiving informed consent) should be able to reliably learn about the sensitive information that has been collected on each individual. This idea is broadly referred to as privacy. Our concern is in the study of privacy as it relates to precision medicine. Precision medicine

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.  
© 2024 The Authors. Stat published by John Wiley & Sons Ltd.

What can we do?

**Differential Privacy:** A rigorous mathematical standard for privacy, that, if achieved, provides provable guarantees for individual protection.

## Family Income

---

1 60,000

2 58,000

3 63,000

4 95,000

5 61,000

6 12,000

7 63,000

8 125,000

9 57,000

10 55,000

---

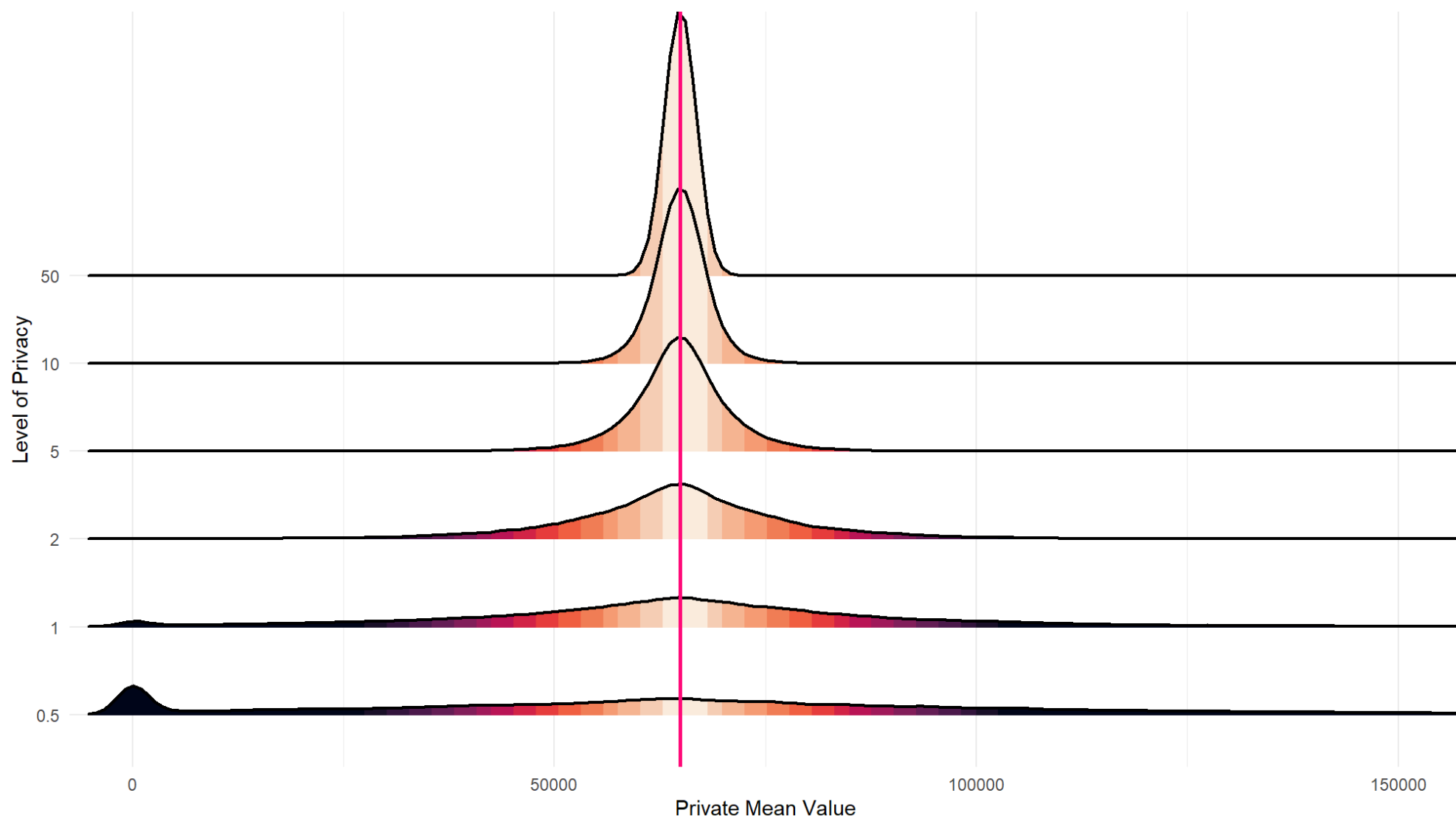
$$\text{Mean} = \frac{60 + 58 + 63 + 95 + 61 + 12 + 63 + 125 + 57}{10} = 6$$



$$\text{Private Mean} = \text{Mean} + \text{Random Noise}$$

- 3994.80 produces:  $64,900 + 3994.80 = 68,894.80$
- - 4987.57 produces:  $64,900 - 4987.57 = 59,912.43$
- - 12943.21 produces:  $64,900 - 12,943.21 = 51,956.79$
- 35374.81 produces:  $64,900 + 35,374.81 = 100,274.81$

The specific value of the private mean depends on the random noise. It may be close or far from the truth.



ORIGINAL ARTICLE

Differentially private outcome-weighted learning for dynamic treatment regime estimation

Dylan Spicker<sup>1</sup> | Erica E. M. Moodie<sup>2</sup> | Susan M. Shortreed<sup>3,4</sup>

<sup>1</sup>Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, New Brunswick, Canada

<sup>2</sup>Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada

<sup>3</sup>Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA

<sup>4</sup>Department of Biostatistics, University of Washington, Seattle, WA, USA

Correspondence

Dylan Spicker, Department of Mathematics and Statistics, University of New Brunswick (Saint John), Saint John, NB, Canada.  
Email: dylan.spicker@unb.ca

Funding Information

Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number R01 MH114873. This research was undertaken, in part, thanks to funding from the Canada Excellence Research Chairs Program. Additionally, Dylan Spicker was supported by the Canadian Statistical Sciences Institute (CANSSI) and STATLAB and Erica E.M. Moodie by a chercheur de mérite career award from the FRQ, Santé. The STAR\*D study was supported by NIMH Contract #N01MH90003 to the University of Texas Southwestern Medical Center. The ClinicalTrials.gov identifier is CT00021528.

Precision medicine is a framework for developing recommendations that seeks to determine the optimal sequence of treatments, tailored to all of the relevant, observable patient-level characteristics. Because precision medicine relies on highly sensitive, patient-level data, ensuring the privacy of participants is of great importance. Dynamic treatment regimes (DTRs) provide one formalization of precision medicine in a longitudinal setting. Outcome-weighted learning (OWL) is a family of techniques for estimating optimal DTRs based on observational data. OWL techniques leverage support vector machines (SVMs) classifiers in order to perform estimation. SVMs perform classification based on a set of influential points in the data known as support vectors. The classification rule produced by SVMs often requires direct access to the support vectors. This, in turn, means that a policy estimated with OWL requires the release of patient-level data for a subset of the sample. As a result, the classification rules from SVMs, including those used in OWL, require direct access to the support vectors. This privacy violation is a major concern, particularly in light of the potentially highly sensitive medical data that are used in DTR estimation. In this paper, we propose a mathematical framework for ensuring the privacy of patient-level data, while providing guarantees on the likelihood that individual data points are used as support vectors. We provide the first investigation of DTRs and provide a differentially private OWL estimator, with theoretical results allowing us to quantify the cost of privacy in terms of the accuracy of the private estimators.

KEYWORDS

differential privacy, dynamic treatment regimes, individualized treatment rules, precision medicine, support vector machines

1 | INTRODUCTION

Health data are, by their nature, sensitive information. We assume that patients expect that their sensitive information will remain protected and secure once it has been collected. The expectation is that no one outside of those who were explicitly granted access (such as the researchers receiving informed consent) should be able to reliably learn about the sensitive information that has been collected on each individual. This idea is broadly referred to as privacy. Our concern is in the study of privacy as it relates to precision medicine. Precision medicine

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.  
© 2024 The Authors. Stat published by John Wiley & Sons Ltd.

# Private Personalized Medicine

Treatment is **more impactful** for some patients than others.

Our method achieves **high accuracy** for those patients and provides **privacy** for all patients.

## Pillar #1

Learning from data is crucial for our understanding of the world.

## Pillar #2

Protecting individual privacy is necessary as data become more abundant.

There is an inherent trade-off between these pillars. Privacy research makes this explicit.

Thank you.

dylanspicker.com | dylan.spicker@unb.ca