

The Thin Line Between Personal and and Anonymous Data

Legal boundaries & Statistical realities

lygature



No standard masterclass







Bart Torensma



Daniel Groos



Jente Houweling



Jan van der Laan

Join at [menti.com](https://www.menti.com) | use code 8899 8282

 Mentimeter

Instructions

Go to

www.menti.com

Enter the code

8899 8282



Or use QR code



Join at menti.com | use code 8899 8282

 Mentimeter

I have experience with anonymisation

▶ Start Menti

Yes

No



Join at menti.com | use code 8899 8282

 Mentimeter

I am familiar with statistical disclosure control


▶ Start Menti

Yes

No



Join at menti.com | use code 8899 8282

 Mentimeter

Pseudonymisation and anonymisation have the same meaning

▶ Start Menti

Yes

No



Join at menti.com | use code 8899 8282

 Mentimeter

Anonymous data is useless

▶ Start Menti

Yes

No



Join at menti.com | use code 8899 8282

 Mentimeter

Anonymising health data is (virtually) impossible

▶ Start Menti

Yes

No



Join at menti.com | use code 8899 8282

 Mentimeter

There are no legal, ethical or societal limits to the use of anonymous data

▶ Start Menti

Yes

No



Introduction

Data processing in health research:
what should researchers consider?

- Expectations of citizens & patients: safe and careful processing of their personal data
- Legal framework: need for a legal basis to process personal data

Legal basis often causes issues

- Especially with 'consent'
- Narrow definition of consent since GDPR

Anonymisation a solution?

- GDPR no longer cause for concern
- Privacy concerns addressed

Issues remain...

- Anonymous data less useful
- How to determine when data is anonymous?

How to determine the anonymity of data?



GDPR, recital 26

“The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”



GDPR, recital 26

*“The principles of data protection should apply to **any information concerning an identified or identifiable natural person**. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”*



GDPR, recital 26

*“The principles of data protection should apply to any information concerning an identified or identifiable natural person. **Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person.** To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”*



GDPR, recital 26

*“The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. **To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.** The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.”*



GDPR, recital 26

*“The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. **The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.**”*

How to determine anonymity?

Discussion among legal experts on how to assess whether data is anonymous

- Given all the circumstances under which data is processed, is reidentification:
 - No longer reasonably possible ?
 - No longer possible in absolute sense?

Not only a legal issue

- Struggle for researchers wishing to share or publish data
- Need for legal, technical and statistical support in assessing risk of reidentification

How to determine risk of reidentification in practice?

Statistical Disclosure Control in practice practice

RIVM strives to make data openly available to the greatest extent possible

As a (scientific) research institute



Open Science
Open Access

As a public organisation



'Wet Open Overheid'
Transparency

Importance of anonymity at RIVM

1. Disclosure of personal data
2. Maintaining appropriate ethical and legal standards (GDPR)
3. Reputation and trust





Legal question -> Statistical question



- Are there variables that, alone or in combination, could lead to identifying individuals?
- Do not strive for complete anonymity, but for an acceptable risk
- Consider the data context, including:
 - Nature and number of sensitive data
 - How the data are shared (under contract or open; controlled or open access)
 - Time since data collection
 - Population size

Statistical Disclosure Control (SDC)

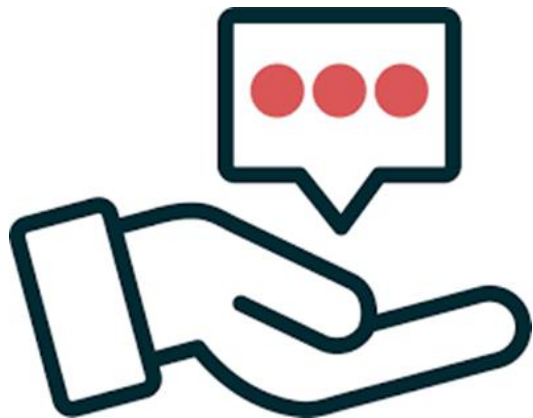
= Determining the risk of disclosure, and
applying measures to minimize
the risk of disclosure



(CBS is an authority in this area in the Netherlands)

SDC group at RIVM

Supports RIVM researchers and data owners on issues regarding disclosure/anonymization for publishing or sharing data.



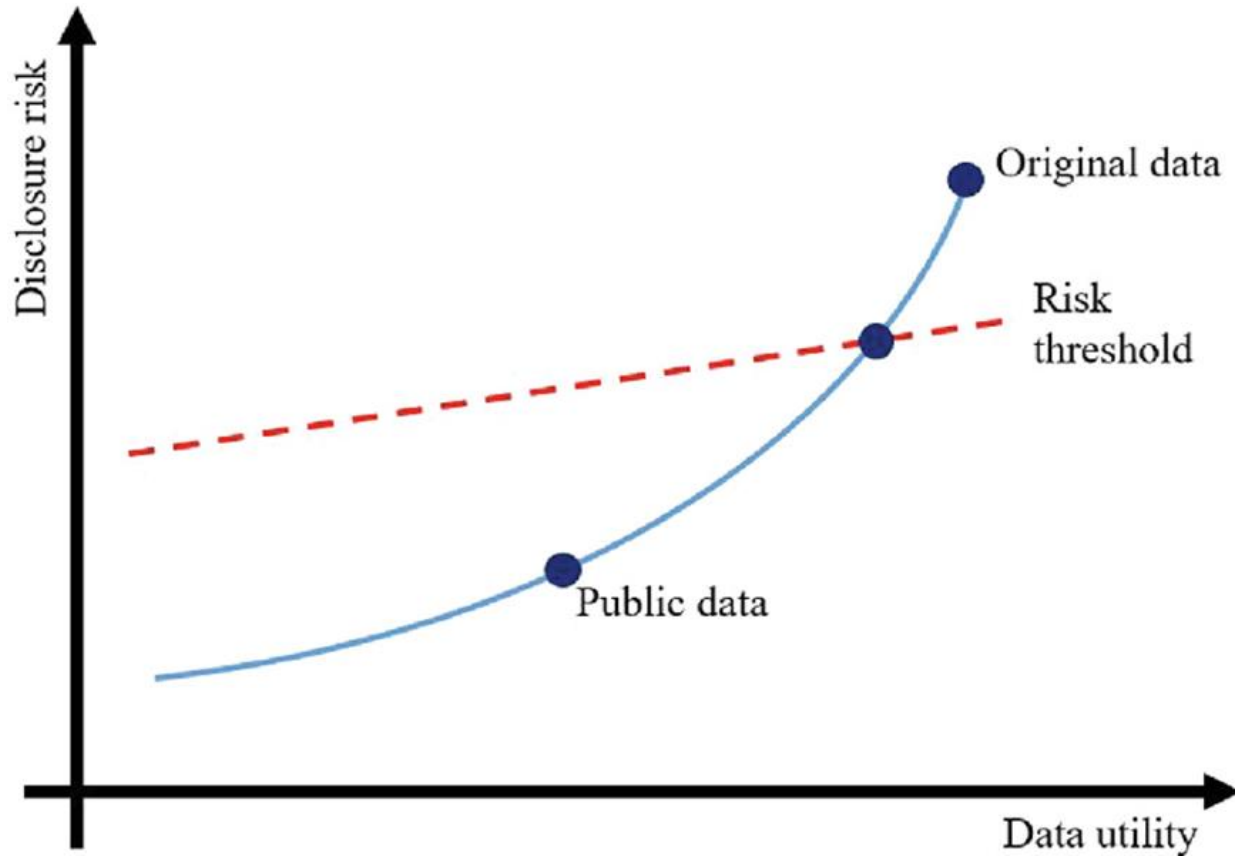
Advice on request



The composition of this group is multi-disciplinary:

- Data stewards
- Data managers
- Statisticians
- Privacy coordinators
- Legal experts
- CDO

The basic principles of SDC

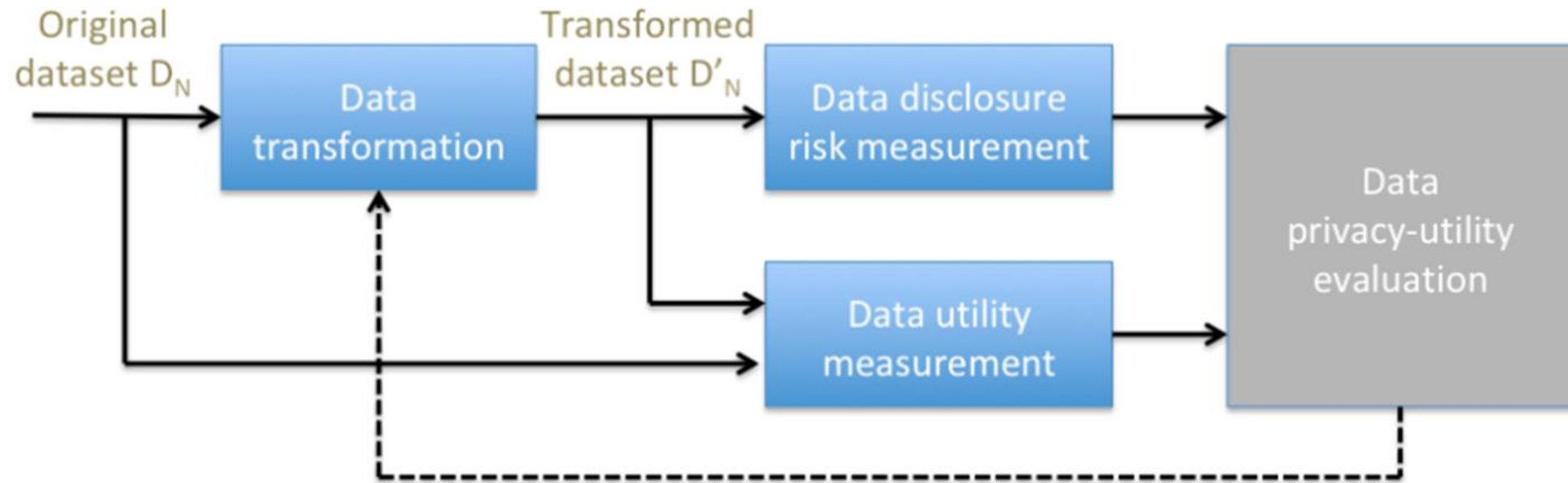


- Disclosure occurs when a person/organisation (the intruder) uses published data to find and reveal sensitive and/or unknown information about a data subject.
- The main goal of SDC is to minimise potential risk of disclosure to an acceptable level while sharing as much data as possible.
- The basic principle of SDC is that it is impossible to reduce the probability of re-identification to zero, so instead one needs to control the risk of disclosure.

Adapted from Magder et al. UKDS, 2021

A generic model of data anonymisation process

Information on context, usage and
and data environment



Is it safe to publish this dataset?

Patient number	Date of birth	Residence	Date of diagnosis
800531	12-09-1963	Utrecht	01-05-2024
800532	03-02-1968	Zeist	05-04-2024
800533	21-12-1978	Arnhem	10-02-2024
800534	11-04-1970	Arnhem	10-02-2024
800535	09-04-1965	Amsterdam	22-02-2024
...

Is it safe to publish this dataset?

Recoding

Patient number	Date of birth	Residence	Date of diagnosis
374619	60	Utrecht	May-2024
883761	56	Utrecht	April-2024
492987	45	Gelderland	February-2024
278114	54	Gelderland	February-2024
981745	59	Noord-Holland	February-2024
...

Is it safe to publish this dataset?

Recoding

Patient number	Date of birth	Residence	Date of diagnosis
374619	60	Utrecht	May-2024
883761	56	Utrecht	April-2024
492987	45	Gelderland	February-2024
278114	54	Gelderland	February-2024
981745	59	Zuid-Holland	February-2024
...
728559	110	Zuid-Holland	NA
...

Lenie M.
(Rotterdam)

Anonymity of data: legal analysis

Legal debate on the boundary between personal and anonymous data data

- Reidentification no longer *reasonably* possible?
- Reidentification no longer possible in *absolute* sense?

Difference of approach

- Relative or Contextual approach: who has access under what circumstances?
- Absolute approach: nature of the data is decisive

Anonymity of data: legal analysis

Relative vs absolute approach

- Opinion 05/2014 on Anonymisation Techniques
 - Singling out, linking, inferring
- Court of Justice of the European Union
 - **Patrick Breyer v Bundesrepublik Deutschland**
 - Dynamic IP address personal data?
 - ‘Means likely reasonably to be used’ ≠ means prohibited by law, disproportionate effort (time, cost, manpower)
 - Relevant *who* has access

Anonymity of data: legal analysis

Relative vs absolute approach

- Court of Justice of the European Union
 - **SRB v EDPS**
 - SRB invited shareholders to submit comments
 - Coded comments shared with third party
 - Data held by third party personal data?
 - CJEU: “not necessarily...”
 - Always necessary to assess whether data recipient reasonably able to re-identify data subjects

Anonymity of data: legal analysis

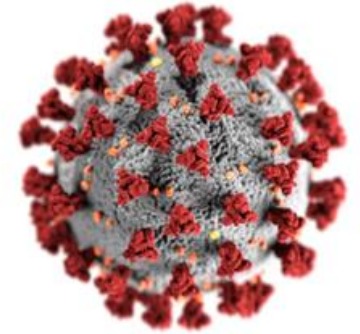
Relative vs absolute approach: operationalisation?

- Absolute approach: focus on techniques (singling out, linking, inferring)
- Operationalising relative approach less straightforward...

Legal boundaries, statistical realities

- Boundary between personal and anonymous
- Legal analysis → abstraction of statistical reality

Case: Study on behavioural measures and well-being during COVID-19 pandemic



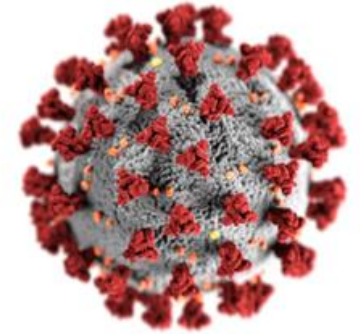
Objectives:

To assess public perception of the behavioural measures and recommendations, their impact on personal well-being, and whether people were complying

Topics:

- Compliance with corona measures
- Behaviour
- Well-being and lifestyle
- Public support for corona measures
- Government communication and trust
- Vaccination willingness

Case: Study on behavioural measures and well-being during COVID-19 pandemic



Characteristics study:

- Longitudinal survey-based study
- April 2020 to September 2022 (21 waves)
- N= 189.619 unique respondents
- >2000 variables

Characteristics data:

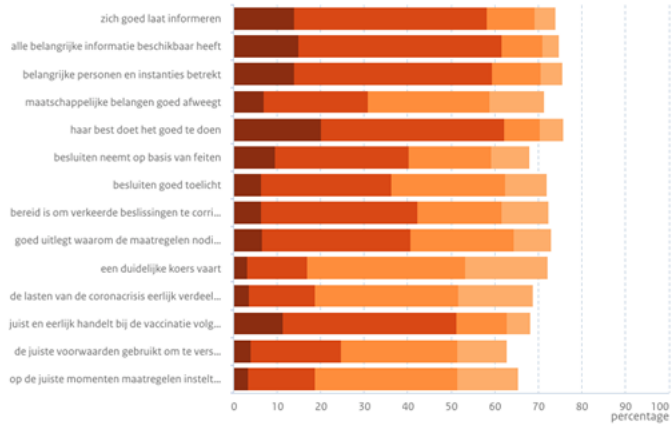
- No sampling weights
- No direct identifiers; several indirect identifiers
- Sensitive information

Case study:

- wave 1 to 14
- N= 175.247 unique respondents (~1% of dutch population)
- >1500 variables

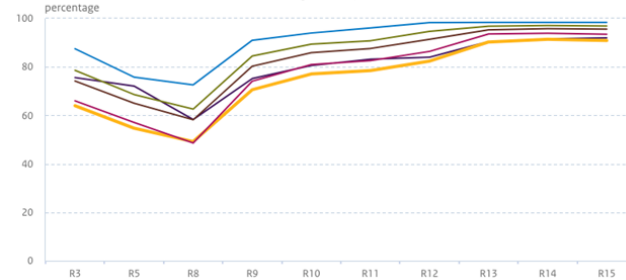
Bij het bepalen van de maatregelen denk ik dat de Nederlandse overheid

Meting 15, 8 - 12 september



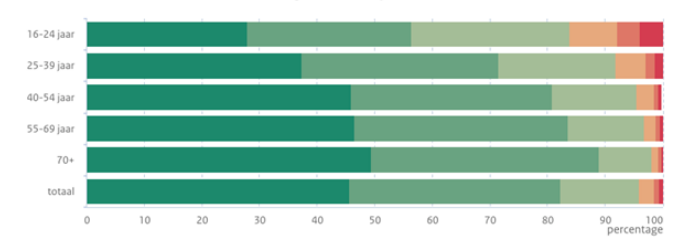
Verandering vaccinatiebereidheid naar leeftijd

Meting 3, 5, 8 t/m 15



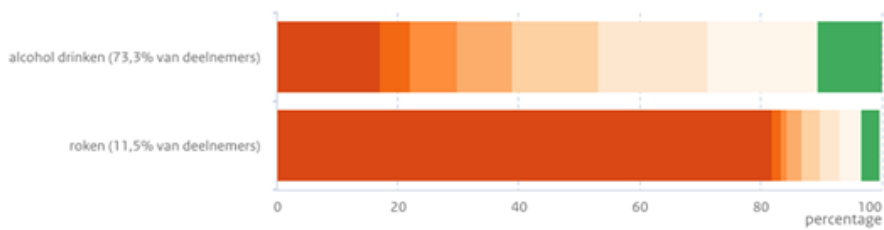
Mentaal welbevinden (angstig) naar leeftijd

Meting 15, 8 - 12 september



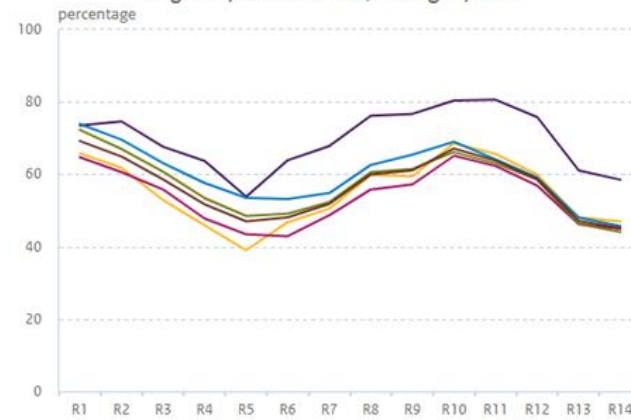
Middelengebruik in de afgelopen 7 dagen

Meting 15, 8 - 12 september



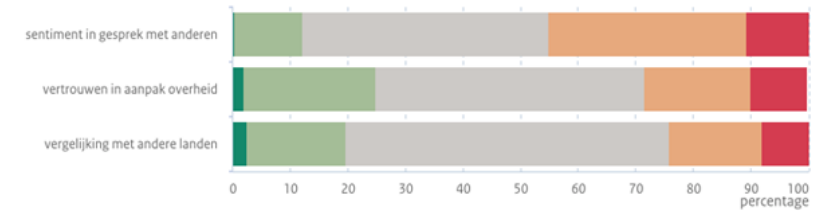
Verandering in eenzaamheid naar leeftijd

Enigszins /sterk eenzaam; Meting 1 t/m 14



Beeld van de aanpak van het coronavirus door de Nederlandse overheid

Meting 15, 8 - 12 september



Selecting key variables

Key variables – combinations of variables (indirect identifiers) that when taken together can identify a respondent, e.g. education, age, employment, religious affiliations, household size, geographic area.

Here, potential key variables are:

"geslacht", "leeftijd_cat7", "leeftijd_cat16",
"gemeente", "stadsdeel", "Gemnum", "GGD_TOT",
"opleiding", "geboorteland",
"woonsituatie", "woontalleen", "werksituatie"

Selecting key variables

Key variables – combinations of variables (indirect identifiers) that when taken together can identify a respondent, e.g. education, age, employment, religious affiliations, household size, geographic area.

Here, potential key variables are:

"**geslacht**", "**leeftijd_cat7**", "leeftijd_cat16",
"gemeente", "stadsdeel", "Gemnum", "**GGD_TOT**",
"**opleiding**", "**geboorteland**",
"woonsituatie", "**woontalleen**", "werksituatie"

Using R-package sdcMicro to estimate disclosure risk in the dataset

SUDA scores: how much does each
variable contribute to the risk

variable	contribution
geslacht	29.55
leeftijd_cat7	62.96
GGD_TOT	91.67
opleiding	74.58
geboorteland	79.51
woontalleen	27.11

The loaded dataset consists of 175247 records and 32 variables.

Categorical key variable(s): geslacht leeftijd_cat7 GGD_TOT opleiding geboorteland woontalleen

Computation time

The current computation time was ~ 6.29 seconds .

Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
geslacht	3 (3)	87449.500 (87449.500)	59958 (59958)
leeftijd_cat7	7 (7)	21905.875 (21905.875)	967 (967)
GGD_TOT	26 (26)	6490.630 (6490.630)	331 (331)
opleiding	10 (10)	19441.778 (19441.778)	394 (394)
geboorteland	10 (10)	19427.444 (19427.444)	120 (120)
woontalleen	2 (2)	87623.500 (87623.500)	27377 (27377)

Risk measures for categorical key variables

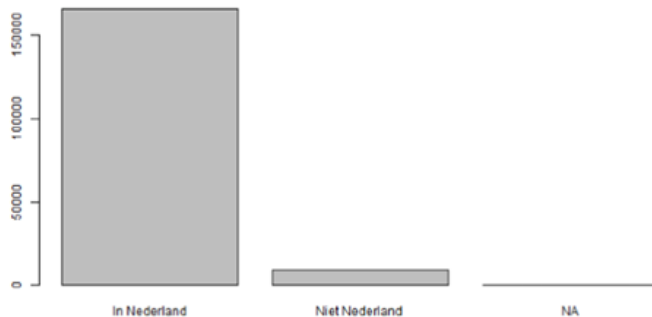
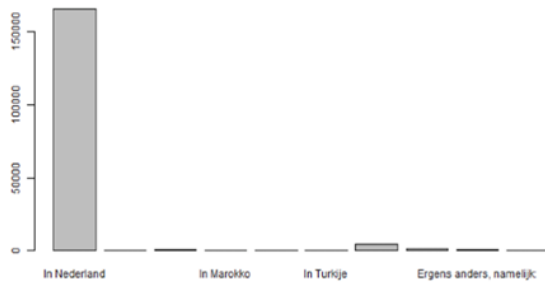
We expect 7114.29 (4.06%) re-identifications in the population, as compared to 7114.29 (4.06%) re-identifications in the original data.

13275 observations have a higher risk than the risk in the main part of the data, as compared to 13275 observations in the original data. 

Global risk: average of individual risk scores
(probability that any individual in the data set can
be re-identified)

Recoding

Changing 'geboorteland' from 10 to 2 categories: 'Nederland' and 'Niet Nederland'
Effect: from **4.06%** risk to **3.21%** risk.
This might seem neglectable, but it is actuality a 21% decrease.



Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
geslacht	3 (3)	87449.500 (87449.500)	59958 (59958)
leeftijd_cat7	7 (7)	21905.875 (21905.875)	967 (967)
GGD_TOT	26 (26)	6490.630 (6490.630)	331 (331)
opleiding	10 (10)	19441.778 (19441.778)	394 (394)
geboorteland	3 (10)	87423.500 (19427.444)	8773 (120)
woontalleen	2 (2)	87623.500 (87623.500)	27377 (27377)

Risk measures for categorical key variables

We expect **5623.29** (**3.21%**) re-identifications in the population, as compared to **7114.29** (**4.06%**) re-identifications in the original data.

10723 observations have a higher risk than the risk in the main part of the data, as compared to **13275** observations in the original data. ⓘ

Violations of K-anonymity

2209 (1.295%) observations violate 2-anonymity (= sample uniques/ have a unique key)

4161 (2.374%) observations violate 3-anonymity

To achieve 3-anonymity, a dataset should be modified in such a way that each combination of attributes (features) shared by at least three different individuals appears in the dataset multiple times

Effect k-anonymisation on disclosure risk

Risk measures

2570 observations have a higher risk than the risk in the main part of the data, as compared to 13275 observations in the original data ⓘ

Based on the individual re-identification risk, we expect 2823.2 re-identifications (1.61%) in the anonymized data set. In the original dataset we expected 7114.29 (4.06%) re-identifications.

1st Step: Recoding

4.06% to **3.21%**

2nd Step: k-anonymisation

3.21% to **1.61%**

Effect k-anonymisation on data utility

How many values for each variable have been suppressed (replaced by NA)?

Do we find this information loss acceptable?

Information on k-anonymity

Below the number of observations violating k-anonymity is shown for the original data and the modified dataset

k-anonymity	Modified data	Original data
2-anonymity	0 (0.000%)	2269 (1.295%)
3-anonymity	0 (0.000%)	4161 (2.374%)
5-anonymity	201 (0.115%)	7323 (4.179%)

Information on local suppression

Below the number of suppressions (values set to a missing value (NA)) due to the last run of the local suppression algorithm. The table also displays the number of missing values (NA) per variable before applying the local suppression algorithm as well as the total number of missing values in each variable after applying local suppression (sum of initial missings and suppressions).

Key variable	Number of suppressions	Total missing values (NA) before applying local suppression	Total missing values (NA) after applying local suppression
geslacht	0 (0.000%)	348 (0.199%)	348 (0.199%)
leeftijd_cat7	3 (0.002%)	0 (0.000%)	3 (0.002%)
GGD_TOT	2499 (1.426%)	0 (0.000%)	2499 (1.426%)
opleiding	36 (0.021%)	271 (0.155%)	307 (0.175%)
geboorteland	0 (0.000%)	400 (0.228%)	400 (0.228%)
woontalleen	0 (0.000%)	0 (0.000%)	0 (0.000%)

Changing the SDC problem

Only leave out the geographical variable 'GGD_TOT' and apply same anonymisation methods.

Result:

1st Step: Recoding
0.42% to **0.21%** risk

2nd Step: K-Anonymisation

0.21% to **0.19%** risk -> Not worth it!

Summary of dataset and variable selection

The loaded dataset consists of **175247** records and **32** variables.

Categorical key variable(s): **geslacht leeftijd_cat7 opleiding geboorteland woontalleen**

Computation time

The current computation time was ~ **3.07 seconds** .


Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level (with frequency >0) for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

Variable name	Number of levels	Average frequency	Frequency of smallest level (>0)
geslacht	3 (3)	87449.500 (87449.500)	59958 (59958)
leeftijd_cat7	7 (7)	21905.875 (21905.875)	967 (967)
opleiding	10 (10)	19441.778 (19441.778)	394 (394)
geboorteland	10 (10)	19427.444 (19427.444)	120 (120)
woontalleen	2 (2)	87623.500 (87623.500)	27377 (27377)

Risk measures for categorical key variables

We expect **741.92** (**0.42%**) re-identifications in the population, as compared to **741.92** (**0.42%**) re-identifications in the original data.

1274 observations have a higher risk than the risk in the main part of the data, as compared to **1274** observations in the original data. 

Conclusion of case study

Omitting geographical variable 'GGD_TOT' → too much information loss (data utility)

So, after reducing the risk by:

1. Recoding the variable Country-of-birth
2. Omitting variable Municipality
3. Local suppression to achieve 3-anonymity
4. Replacing variable Respnum by random RespID
5. Sorting the dataset by this random RespID

the dataset was shared with one party 'under contract'.



Take-home message

Absolute anonymity generally unattainable. Minimizing the risk of disclosure is key

- Addressed through Statistical Disclosure Control

Discussion on absolute vs relative anonymity ongoing

Does the European Health Data Space (EHDS) acknowledge the relative approach?

- *“A certain risk is assumed with electronic health data that can remain particularly sensitive even when anonymized”* (Recital 64)
- *“There remains a residual risk that the capacity to reidentify could be or becomes available, **beyond the means reasonably likely to be used**”*

Take-home message

Importance of role of background information.

Essential to properly map data environment, considering factors such as:

- Time since data collection
- Sample size
- How the data is shared (under contract or openly)
- The sensitivity of the data

Special attention should be given to geographic variables, as their impact on disclosure risk is usually high

Take-home message

Anonymity: not about escaping the GDPR

- Anonymisation form of data processing → should be compatible with original purpose
- If consent original legal basis → secondary use of anonymised data should not be incompatible with original consent
- Research with anonymised data should remain within reasonable expectations of data subjects
- Ethical standards still apply
- Hard to build trust, easy to lose

Thank you for your attention

Bart Torensma

bart@toensmaresearch.nl

Daniel Groos

daniel.groos@lygature.org

Jente Houweling

jente.houweling@rivm.nl

Jan van der Laan

jan.van.der.laan@rivm.nl

The logo for Lygature, featuring the word "lygature" in a white, lowercase, sans-serif font. The letter "y" is stylized with a thin orange vertical line through its center. The text is set against a dark teal rectangular background.The logo for Coreon, consisting of a green stylized icon of a person's head and shoulders to the left of the word "coreon" in a bold, green, lowercase sans-serif font. Below "coreon" is the text "Commissie Regelgeving Onderzoek" in a smaller, green, lowercase sans-serif font.