

# Methods to assess and mitigate re-identification risks when sharing research data

Gustav Nilsson, 2021-11-11



**SND**

Svensk nationell datatjänst | Göteborgs universitet - Karolinska institutet - Lunds universitet - Umeå universitet - Stockholms universitet - Sveriges lantbruksuniversitet - Uppsala universitet



# Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

Mandeep R Mehra, Sapan S Desai, Frank Ruschitzka, Amit N Patel

### Summary

**Background** Hydroxychloroquine or chloroquine, often in combination with a second-generation macrolide, are broadly used for treatment of COVID-19, despite no conclusive evidence of their benefit. Although generally safe when used for approved indications such as autoimmune disease or malaria, the safety and benefit of these treatment regimens are poorly evaluated in COVID-19.

**Methods** We did a multinational registry analysis of the use of hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19. The registry comprised data from 671 hospitals in 34 continents. We included patients hospitalised between Dec 20, 2019, and April 14, 2020, with a positive laboratory finding for SARS-CoV-2. Patients who received one of the treatments of interest within 48 h of diagnosis were included in one of four treatment groups (chloroquine alone, chloroquine with a macrolide, hydroxychloroquine alone, or hydroxychloroquine with a macrolide), and patients who received none of these treatments formed the control group. Patients for whom one of the treatments of interest was initiated more than 48 h after diagnosis or while they were on mechanical ventilation, as well as patients who received remdesivir, were excluded. The main outcomes of interest were in-hospital mortality and the occurrence of de-novo ventricular arrhythmias (as defined or sustained ventricular tachycardia or ventricular fibrillation).

**Findings** 96 032 patients (mean age 53·8 years, 46·3% women) with COVID-19 were hospitalised during the study period and met the inclusion criteria. Of these patients, 46 221 were in the treatment groups (1868 received chloroquine, 3783 received chloroquine with a macrolide, 3016 received hydroxychloroquine, and 6221 received hydroxychloroquine with a macrolide) and 49 811 patients were in the control group. 10 698 (11·1%) patients died in hospital. After controlling for multiple confounding factors (eg, sex, race or ethnicity, body-mass index, underlying cardiovascular disease and its risk factors, diabetes, underlying lung disease, smoking, immunosuppressed condition, and baseline disease severity), when compared with mortality in the control group (9·3%), hydroxychloroquine (18·0%; hazard ratio 1·335, 95% CI 1·223–1·457), hydroxychloroquine with a macrolide (23·8%; 1·447, 1·368–1·531), chloroquine (16·4%; 1·365, 1·218–1·531), and chloroquine with a macrolide (22·2%; 1·368, 1·273–1·469) were each independently associated with an increased risk of in-hospital mortality. Compared with the control group (0·3%), hydroxychloroquine (6·5%; 2·367, 1·935–2·900), hydroxychloroquine with a macrolide (8·1%; 5·106, 4·106–5·983), chloroquine (4·3%; 1·751, 1·240–4·596), and chloroquine with a macrolide (6·5%; 4·011, 3·344–4·812) were independently associated with an increased risk of de-novo ventricular arrhythmia during hospitalisation.

**Interpretation** We were unable to confirm a benefit of hydroxychloroquine or chloroquine, when used alone or with a macrolide, on in-hospital outcomes for COVID-19. Each of these drug regimens was associated with decreased in-hospital mortality, but also with an increased frequency of ventricular arrhythmias when used for treatment of COVID-19.

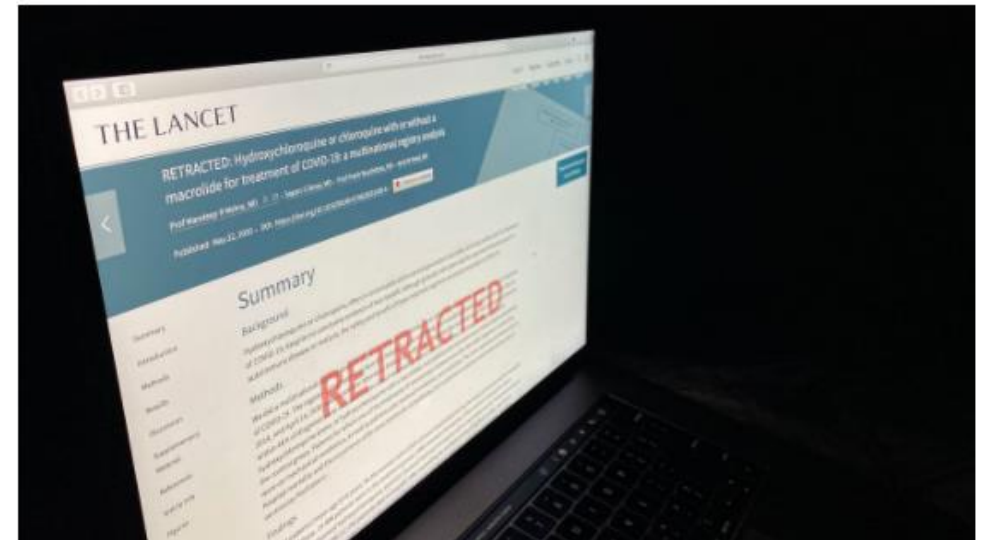
**Funding** William P. Gray Distinguished Chair in Advanced Cardiovascular Medicine at Brigham and Women's Hospital.

Copyright © 2020 Elsevier Ltd. All rights reserved.



[https://doi.org/10.1016/S0140-6736\(20\)31180-6](https://doi.org/10.1016/S0140-6736(20)31180-6)

### SHARE




E. PETERSEN/SCIENCE

## Who's to blame? These three scientists are at the heart of the Surgisphere COVID-19 scandal

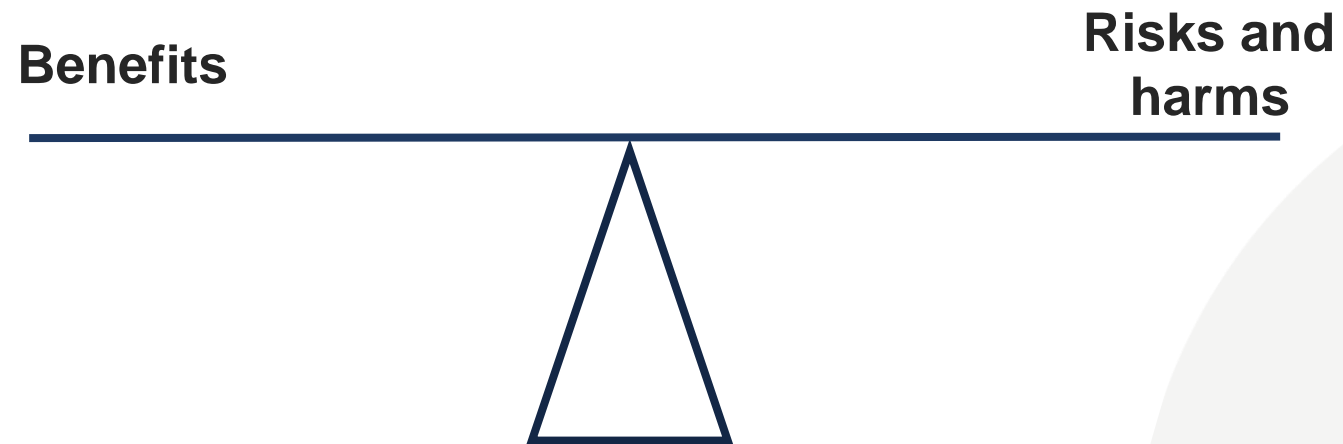
By Charles Piller | Jun. 8, 2020, 7:00 PM




# Data sharing allows researchers to:

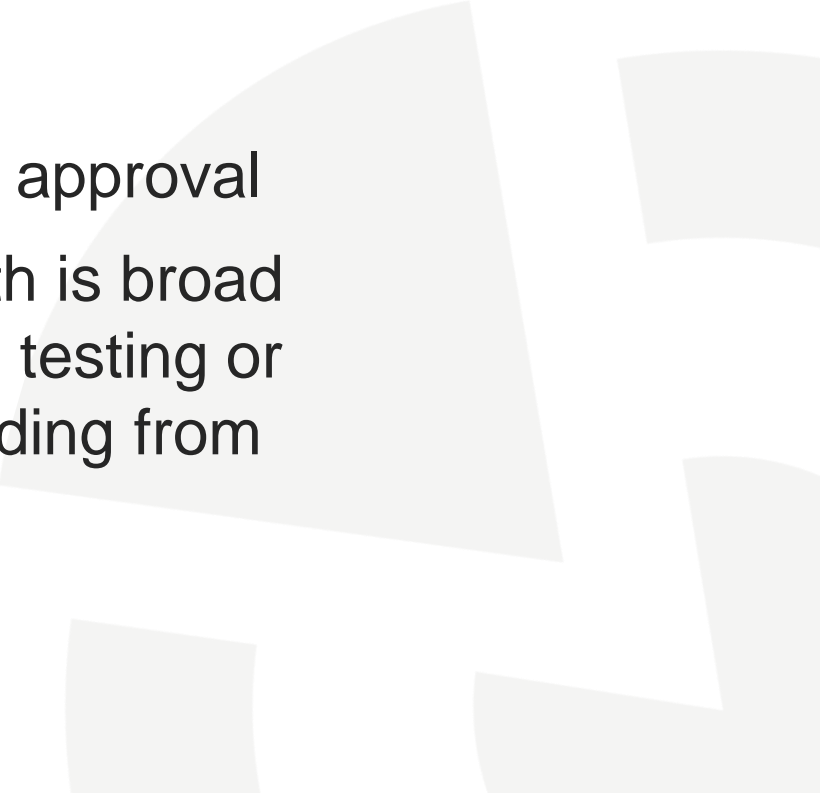
- Verify that registry data exist
  - Reproduce reported results
  - Reanalyse data with different methods, assumptions etc
  - Plan new analyses, e.g. sample size planning/power calculation
  - Validate new results in independent datasets
- 

# Legal and ethical issues in data sharing






# GDPR (General Data Protection Regulation)

- Personal data are data that can be attributed to a living natural person by means reasonably likely to be used
  - Processing of personal data must have a legal basis
    - Research in the public interest is recommended basis
  - Processing of sensitive personal data requires ethics approval
  - Definition of sensitive personal data concerning health is broad and covers clinical data, information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples, etc.
- 




# Sharing data from humans

- Anonymized data can be shared fully openly
  - Non-anonymized data can be shared through a controlled access model
    - Offered under "SND 2.0"
  - Other options for data that have not been anonymized
    - Share summary statistics, distributions etc.
    - Share correlation matrices
- 



# What is re-identification

- An **attacker** combines **target data** with **reference** data and achieves **matching**; data about individuals can then be inferred
  - Can be **deterministic** or **probabilistic**
- 

# A systematic review on reidentification attacks

- Known reidentification attacks on health data at the time (2011) were mainly academic
- Success rates were very low in data that were anonymized properly

<https://doi.org/10.1371/journal.pone.0126772>

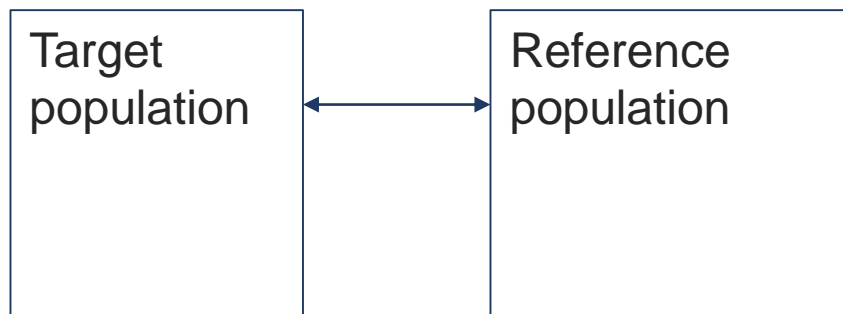
ID	Study	Pub Year <sup>§</sup>	Health data included?	Profession of adversary	Number of individuals re-identified	Country of adversary	Proper de-identification of attacked data ?	Re-identification verified ?
A	[70]	2001	No	Researchers	29 of 273	Germany	"Factually anonymous"	Yes (records containing insurance numbers only)
B	[71]	2001	No	Researchers	75% of 11,000	USA	Direct identifiers removed	No
C	[67]	2002	Yes	Researcher	1 of 135,000	USA	Removal of names and addresses	Yes
	[56]	2003	No	Researchers	219 unique matches, 112 with 2 possibilities, 8 confirmed	UK	Yes	Verified matches, but not identities
D	[22]	2006	No	Journalist	1 of 657,000	USA	No	Yes (with individual)
E	[72]	2006	Yes	Researchers	79% of 550	USA	No	Verified (with original data set)
	[73]	2006	No	Researchers	Of 133 users, 60% of those who mention at least 8 movies	USA	Direct identifiers removed	No
F	[52]	2006	Yes	Expert Witness	18 of 20	USA	Only type of cancer, zip code and date of diagnosis included in request	Yes (verified by the Department of Health)
G	[74]	2007	No	Researchers	2,400 of 4.4 million	USA	Identifying information removed	Verified using original data
	[53]	2007	Yes	Broadcaster	1	Canada	Direct Identifiers removed & possibly other unknown de-id methods used	Yes
H	[23]	2008	No	Researchers	2 of 50	USA	Direct identifiers removed+maybe perturbation	No
I	[75]	2009	Yes	Researcher	1 of 3,510	Canada	Direct identifiers removed	Yes
J	[76]	2009	No	Researchers	30.8% of 150 pairs of nodes	USA	Identifying information removed	Verified using ground-truth mapping of the 2 networks
K	[57,58] <sup>???</sup>	2010	Yes	Researchers	2 of 15,000	USA	Yes - HIPAA Safe Harbor	Yes

(§This is the first year that the report or article appears. Some of the reports we cite have been updated at later dates. Some reports describe re-identification attacks that may have occurred in earlier years. ✂ Since the appearance of the original results in 2010 a second article has been published more recently).  
doi:10.1371/journal.pone.0028071.t002



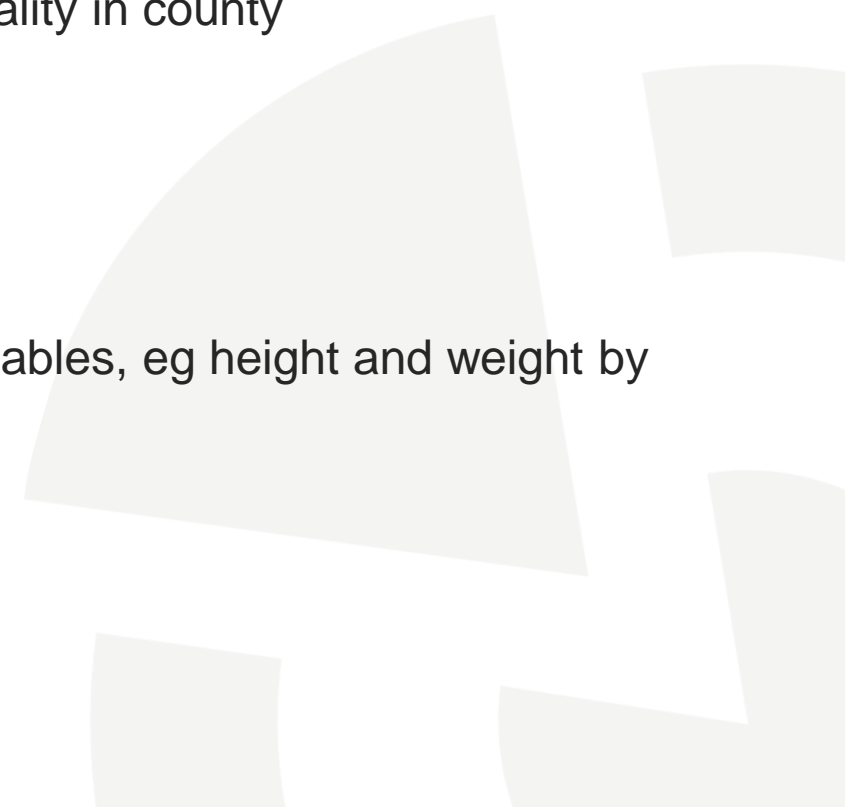
# Assessing risk of reidentification

- Risk depends on **likelihood of reidentification** and **sensitivity of data**
- Threat models
  - Self-identification by participant
  - Targeted reidentification
  - Mass reidentification
- Data uniqueness: can a participant be singled out?
- Is there a reference dataset to which data could be matched?

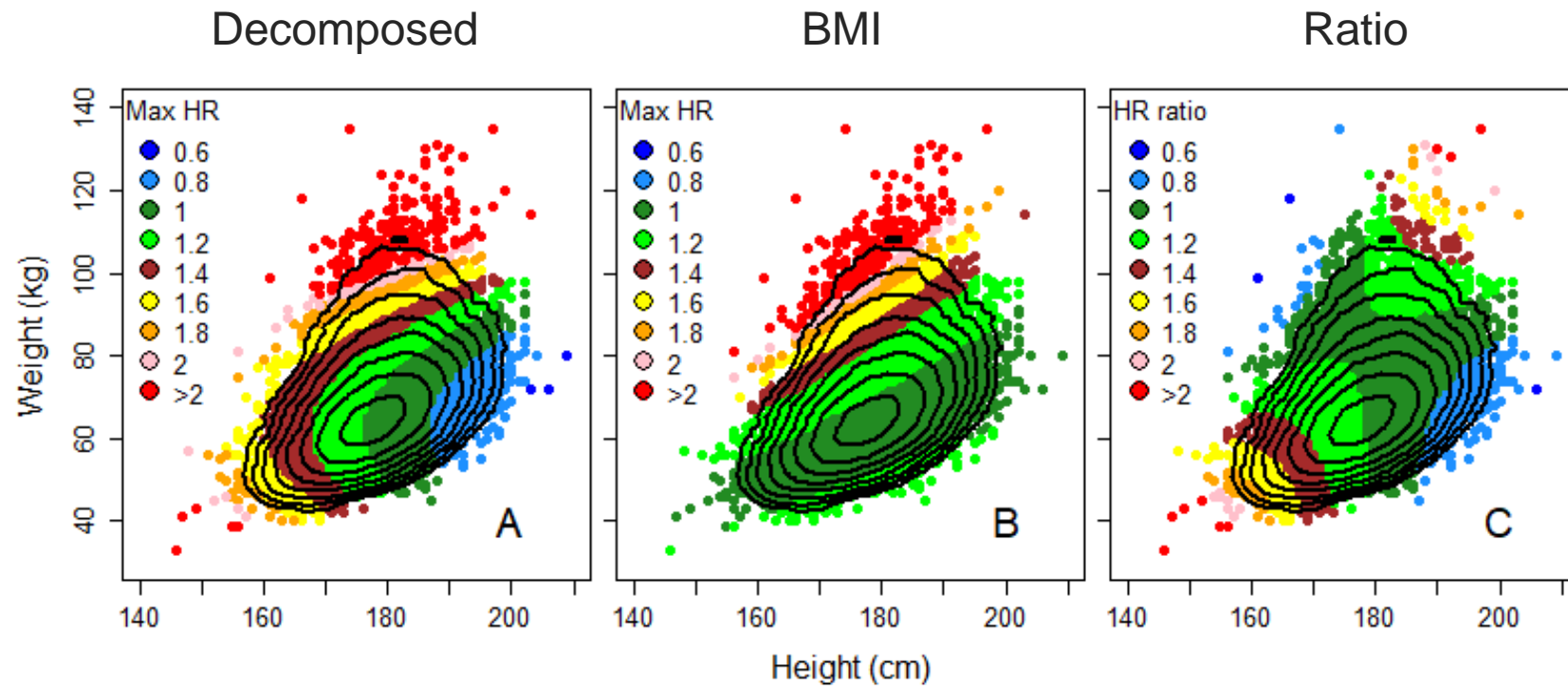




# Actions on variables

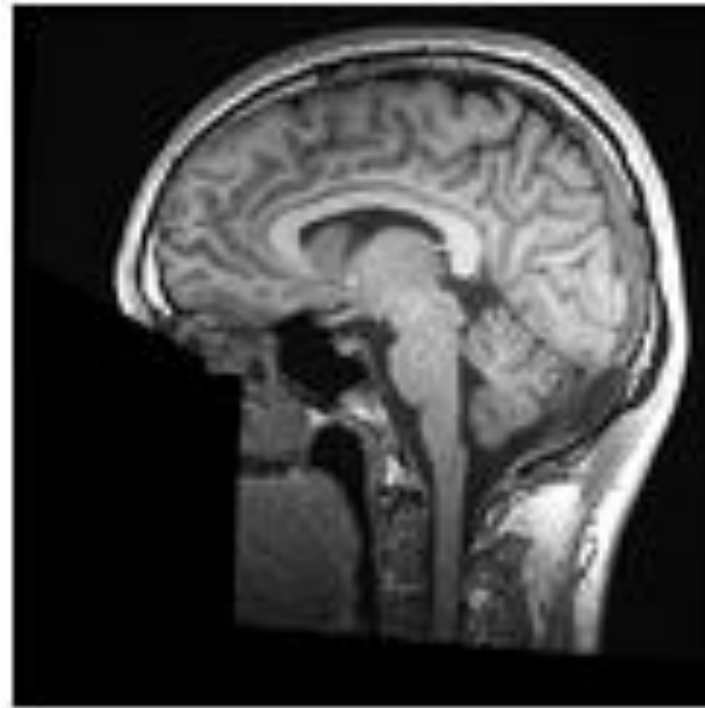
- Aggregation/categorization/binning
    - Continuous variables can be binned, eg age 21-30 years. May require differently sized bins at top or bottom, eg age 81-
    - For hierarchical variables, lower levels can be dropped, eg municipality in county
  - Transformation
    - Preserves internal relations, eg by subtracting a constant
  - Combination
    - Two or more variables can be replaced by a combination of the variables, eg height and weight by BMI
  - Censoring
- 

# Example: singling out in a bivariate distribution



# Example: risk mitigation of biometric data

- Defacing of MRI images of the head



# Example: genetic data

- Should be assessed case-by-case
- Combination of allele frequencies identifying?
- Information about health?
  - APOE genotype – prognostic information not communicated to participants

	A	B	C	D
1	rs4680..G.Val..A.Met.	rs7412	rs429358	ApoE
2	AA	CC	TT	e3/e3
3	AA	CC	CT	e4/e3
4	AG	CC	TT	e3/e3
5	AG	CC	CT	e4/e3
6	AA	CC	TT	e3/e3
7	AA	CC	CT	e4/e3
8	AG	CC	CT	e4/e3
9	GG	CC	TT	e3/e3
10	AG	CC	CT	e4/e3
11	AA	CC	TT	e3/e3
12	AA	CC	TT	e3/e3
13	GG	CT	TT	e3/e2
14	GG	CC	TT	e3/e3

# Example: unlinked variable

## Variable Description

Name CACG1  
Label Satisfied with life at present  
Pre-Question Text And now a few questions about you.  
Question Text At present, how satisfied are you with your LIFE? Very, somewhat, a little, or not at all?  
Dataset [M3\\_MKE2\\_SURVEY\\_N389\\_20180604](#)

Value	Label	Frequency	% of valid	% of all
1	VERY	212	54.64%	54.50%
2	SOMEWHAT	141	36.34%	36.25%
3	A LITTLE	23	5.93%	5.91%
4	NOT AT ALL	12	3.09%	3.08%
7	DON'T KNOW	1		0.26%

Valid	Invalid	Min	Max
388	1	1	4

# Example: data re-use

**TABLE I.** Circadian data and characteristics for serum IL-6 levels measured every 3 hours for 24 hours in healthy men

Subject No.	Age (yr)	Time (clock hr of beginning of sampling interval)							
		19:00	22:00	01:00	04:00	07:00	10:00	13:00	16:00
1	46	1.22	1.38	2.64	2.83*	0.70	0.96	2.41*	1.14
2	47	2.23	4.49*	3.63	4.75*	1.63	1.39	1.90	1.96
3	48	0.81	1.64	2.57*	2.43	2.22	1.74	1.75	0.97
4	50	0.50	1.71	2.65	2.89*	3.18	1.20	2.32	1.10
5†	50	1.00	1.82	2.11*	1.92	1.88	1.12	2.20*	1.35
6	50	0.85	2.72	3.11*	2.64	1.16	1.02	1.26	0.69
7	50	1.71	2.93	5.65*	1.80	3.07	2.45	1.47	1.89
8	58	3.21	10.06	10.18*	4.63	5.01	2.60	3.00	4.38
9	65	2.32	3.18	4.47*	3.36	2.45	1.20	1.39	1.65
10	71	3.74	4.35	5.67	6.83*	4.98	2.08	2.11	2.44
11	72	4.00	3.63	4.40	3.66	5.09*	1.76	2.96	4.24

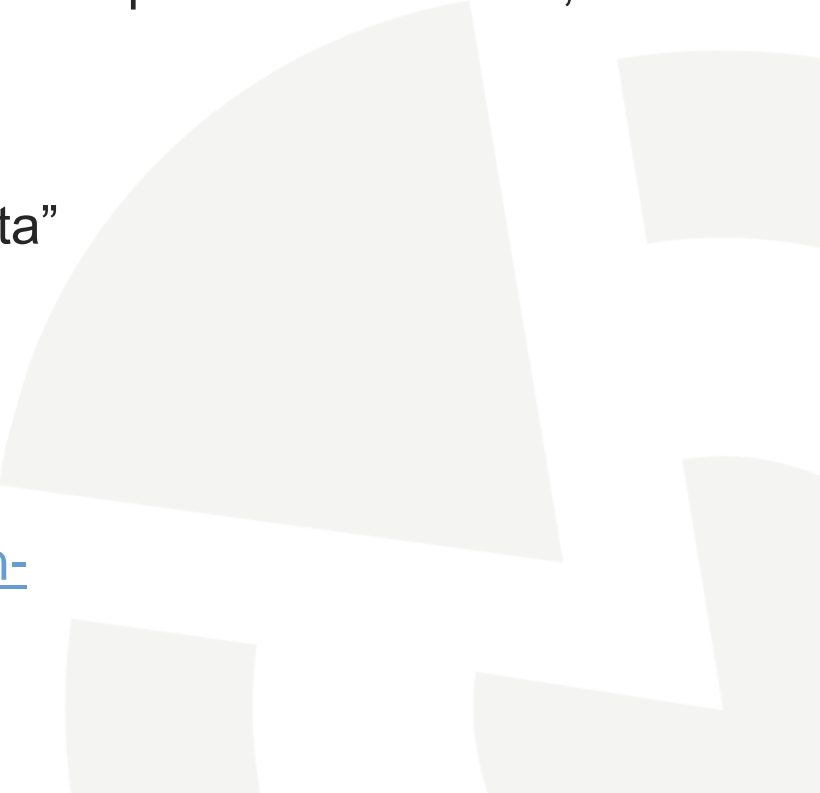


# Recommendations for consent forms when collecting primary data

## Consider phrases like:

- "Data will be published openly in the XX repository [link]"
- "We will remove all data we think could be used to identify you in the published data set, for example name and date of participation"

## Avoid phrases like:

- "No-one outside the research group will have access to your data"
  - "Results will be published only as statistical averages"
  - "Your data will be stored for ten years"
  - Example language in English available e.g. at <https://open-brain-consent.readthedocs.io/en/master/#>
- 





Thank you!





# Further reading

- Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records, Simon et al.  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6450246/>
- Hrynaszkiwicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Bmj*. 2010 Jan 29;340. <https://dx.doi.org/10.1136%2Fbmj.c181>
- Keerie C, Tuck C, Milne G, Eldridge S, Wright N, Lewis SC. Data sharing in clinical trials—practical guidance on anonymising trial datasets. *Trials*. 2018 Dec;19(1):1-8. <https://dx.doi.org/10.1186%2Fs13063-017-2382-9>