THE UNIVERSITY *of* EDINBURGH

Are anonymised datasets from clinical trials truly anonymous?

THE UNIVERSITY *of* EDINBURGH

Aryelly Rodriguez

Principal Supervisor: Professor Steff Lewis

Additional Supervisors: Professor Christopher Weir, Doctor Tracy Jackson and Professor Sandra Eldridge

Submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2024

Declaration

I hereby declare that the work presented in this thesis has been composed solely by myself, has not been submitted in whole or in part for any previous degree application elsewhere, and it is my original work, except where explicitly stated otherwise in the text.

Aryelly Rodriguez

“Success is no accident. It is hard work, perseverance, learning, studying, sacrifice and most of all, love of what you are doing or learning to do.”

—Edson Arantes do Nascimento (Pelé)

Abstract

Background

Funders, regulators and publishers are increasingly requesting that clinical trial researchers share their research data with others, once the primary analysis has been completed. Existing clinical trial data could significantly contribute to expanding medical and scientific knowledge by investigating questions beyond the original study scope, facilitating individual participant data (IPD) meta-analysis, verifying results, and exploring novel methodologies for data analysis.

Anonymisation of IPD before sharing can offer a way to safeguard participants' privacy. While there are several recommendations and guidance available for attempting data anonymisation prior to sharing, completely anonymising data while keeping it usable remains challenging. Moreover, many anonymised datasets are already publicly available for secondary research. However, it remains unclear whether study participants could potentially be at risk of re-identification, and under what circumstances re-identification is more likely to occur.

Methods

In the first phase of this PhD research, a systematic scoping review was conducted to gather publications that reported recommendations on anonymisation for enabling data sharing from clinical trials, to understand what guidance was available to researchers and how publicly available anonymised datasets from clinical trials might have been compiled. Two reviewers, Aryelly Rodriguez with Chris Tuck or Alastair Murray independently assessed titles, abstracts, and full texts for eligibility. One reviewer extracted data from selected papers using thematic synthesis, which was

then reviewed by a second reviewer for accuracy. Results were summarised through narrative analysis.

Moving on to the second phase, I collected a broad selection of publicly available anonymised datasets that have been made available for research purposes extending beyond their original scope, to explore the characteristics of these anonymised datasets, assess the feasibility of applying re-identification risk scores to them, and determine how these scores could be useful. I estimated their re-identification risk scores with three equations designed for calculation of such scores based on the information in the entire dataset. These equations are commonly applied to routinely collected health records and only generate numerical values ranging from 0 (lowest risk) to 1 (maximum risk), without attempting to re-identify individuals within the datasets. Subsequently, I calculated the re-identification risk scores for each dataset, using the three equations. This analysis explored the characteristics of the datasets associated with increased or decreased risk scores, and compared the risk scores to evaluate their practicality for implementation.

In the third and final phase of this PhD research, I used an online exploratory cross-sectional descriptive survey that consisted of both open-ended and closed questions to gather the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets.

Results

The systematic scoping review identified 59 eligible articles (from 43 studies) for inclusion. From these articles, three distinct themes emerged: anonymisation, de-identification and pseudonymisation. The articles also showed that the most commonly recommended anonymisation techniques are removal of direct participant

identifiers, and the careful evaluation and modification of indirect identifiers to minimise the risk of identification. Anonymisation of datasets in conjunction with controlled access was the most recommended method for data sharing.

For the next phase, I contacted data holders and followed their local procedures to access the anonymised datasets. I identified 86 potentially eligible datasets from 18 repositories and successfully secured 76 of them. After full evaluation, 70 datasets met the inclusion criteria and were included in the analysis, representing 14 out of the 18 repositories. Thirty-one datasets were shared with minimal restrictions (open access), while 39 were shared with varying levels of restrictions before access was granted (controlled access). Datasets had, on average, four identifiers and mean risk scores ranging from 0.47 to 0.91. The most common pieces of information present in the datasets that, when combined, may indirectly identify a participant were sex (80%) and age (72.9%).

For the final phase, the exploratory survey had 38 responses to invitation from June 2022 to October 2022. Thirty-five participants (92%) used internal documentation, institutional standard operating procedures and/or published guidance to de-identify/anonymise clinical trials datasets. De-identification followed by anonymisation and then fulfilling data holders' requirements before access was granted (controlled access) was the most common process for releasing the datasets as reported by 18 (47%) participants. Eleven participants (29%) had previous knowledge of re-identification risk estimation but had not used this. Experiences in the process of de-identifying/anonymising the datasets and maintaining such datasets were mostly negative, the main reported issues were lack of resources, guidance, and training.

Conclusions

There is no single standardised set of recommendations on how to anonymise clinical trial datasets for sharing. However, the systematic scoping review showed a developing consensus on techniques used to achieve anonymisation. Researchers in clinical trials still consider that anonymisation techniques by themselves are insufficient to protect participant privacy, and they need to be paired with controlled access.

The second phase of this research confirmed that clinical trial datasets are very rich in personal details and using re-identification risk scores as a measure of this richness is feasible. These scores could inform the anonymisation process of clinical trials datasets to release them for secondary research. We proposed a strategy for incorporating these scores into the decision-making process for releasing clinical trials datasets.

Finally, the majority of responders to the survey reported using documented processes for de-identification and anonymisation. However, our survey results clearly indicate that there are still gaps in the areas of guidance, resources and training to fulfil sharing requests of de-identified/anonymised datasets, and that re-identification risk estimation is an underdeveloped area.

This work will be of interest to the clinical trials research community, funders and publishers seeking to improve the process of anonymisation and foster data sharing.

Plain English summary

Researchers should share their clinical trial data with other investigators after their studies end, as this can help expand medical knowledge. It is important to protect privacy by anonymising this data. But it's tricky to make data completely anonymous while keeping it useful. We wanted to see if it was possible to identify people from data that had been "anonymised".

First, we looked at recommendations for anonymising clinical trial data. A common suggestion is to remove participants' personal details, like name and date of birth. Another suggestion is to be careful with information, such as: sex, race, height and weight. Also, sharing anonymised data with some restrictions and conditions is often recommended.

Next, we checked anonymised datasets available in the public domain, we used three equations to estimate the risk of identifying someone from the data. We found that it is possible to calculate the risk, that some data had more risk than others, and that risk calculation could be informative to researchers.

Then, we asked UK researchers about their experiences with anonymising data. Most used guidelines for this process. They often had to follow strict rules before sharing data.

We found that there isn't one perfect way to make clinical trial data anonymous. Researchers still worry about privacy, even with anonymised data. We also found that there's not enough support for researchers in this area.

Our findings are important for anyone involved in clinical trials. They show the challenges of anonymising data and the need for better support in this area.

Acknowledgements

I extend heartfelt thanks to my supervisors for their invaluable support, my family for their boundless patience, understanding, and encouragement, and to the College of Medicine & Veterinary Medicine at the University of Edinburgh (CMVM/UoE) and Asthma UK Centre for Applied Research for their generous funding and training throughout this endeavour.

Acknowledgements to data providers

Thanks to all data holder/owners for providing access to the data and to all clinical trial participants who permitted the use of their data for secondary research.

Requested acknowledgments statements were:

id	Source (included studies)	Requested acknowledgments statements
1	https://datacompass.lshtm.ac.uk [61] (NCT02104232[62], NCT02111915[63], ISRCTN36436933[64])	No statement required.
2	https://ctu-app.lshtm.ac.uk/freebird [65] (ISRCTN7445979[66], NCT00375258[67], NCT00872469[68] NCT00872469[69], NCT03777488[70])	No statement required
3	https://datashare.is.ed.ac.uk [71] (ISRCTN45178534[72] , ISRCTN25765518[73], IST (Registration not required)[35], ISRCTN71907627[36] ISRCTN89489788[34])	we gratefully acknowledge the IST-3 Collaborative Group, the trial joint sponsors (The University of Edinburgh and the Lothian Health Board), and the chief funding agencies of the study: UK Medical Research Council, Health Foundation UK, Stroke Association UK, Research Council of Norway, Arbetsmarknadens Partners Forsakringsbolag (AFA) Insurances Sweden, Swedish Heart Lung Fund, The Foundation of Marianne and Marcus Wallenberg, Polish Ministry of Science and Education, the Australian Heart Foundation, Australian National Health and Medical Research Council (NHMRC), Swiss National Research Foundation, Swiss Heart Foundation, Assessorato alla Sanita, Regione dell'Umbria, Italy, and Danube University
4	https://www.clinicalstudydatarequest.com [74] (Registration not required[75] , NCT01822899[76] NCT01842607[77], NCT01405053[78], NCT00948766[79])	This publication is based on research using data from the Sponsor companies GlaxoSmithKline Research & Development Ltd, Eisai Limited and Novartis Pharma AG that have been made available to us through secured access. CSDR team or Sponsors have not contributed to or approved, and are not in any way responsible for, the contents of this publication. We thank both Sponsors and CSDR for providing us data and access.
5	http://datadryad.org [80] (ACTRN12616000888460[81], HKCTR-1848[82] Registration no required[83], NCT04523831[84])	No statement required
6	http://yoda.yale.edu [22] (NCT01715285[85], NCT00903331[86] NCT01004432[87], NCT00211133[88])	This study, carried out under YODA Project 2022-4951, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C.. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C..
7	https://www.projectdatasphere.org [21] (NCT00058474[89], NCT00033293[90], NCT00310180[91], NCT00693992[92], NCT00312208[93],NCT00143455[94] NCT00113763[95],NCT00617669[96], NCT00676650[97])	This publication is based on information obtained from https://data.projectdatasphere.org, which is maintained by Project Data Sphere, and includes information that has been made available by the National Cancer Institute and is also available through the National Clinical Trials Network (NCTN)/NCI Community Oncology Research Program (NCORP) Data Archive. The information was collected from the following clinical trials: • A Clinical Trial Comparing Preoperative Radiation Therapy And Capecitabine With or Without Oxaliplatin With Preoperative Radiation Therapy And Continuous Intravenous Infusion Of 5-Fluorouracil With or Without Oxaliplatin In The Treatment Of Patients With Operable Carcinoma Of The Rectum. NCT00058474. • A Pilot Study Randomized Trial of Intravenous Gammaglobulin Therapy for Patients With Neuroblastoma Associated Opsoclonus-Myoclonus-Ataxia Syndrome Treated With Chemotherapy and Prednisone. NCT00033293. • Program for the Assessment of Clinical Cancer Tests (PACCT-1): Trial Assigning Individualized Options for Treatment:The TAILORx Trial. NCT00310180. • Randomized, Phase III, Double-Blind Placebo-Controlled Trial of Sunitinib (NSC #736511) as Maintenance Therapy in Non-progressing Patients Following an Initial Four Cycles of Platinum-Based Combination Chemotherapy in Advanced, Stage IIIB / IV Non-small Cell Lung Cancer. NCT00693992. • A Multicenter Phase III Randomized Trial Comparing Docetaxel in Combination With Doxorubicin and Cyclophosphamide Versus Doxorubicin and Cyclophosphamide Followed by Docetaxel as Adjuvant Treatment of Operable Breast Cancer HER2neu Negative Patients With Positive Axillary Lymph Nodes. NCT00312208. • Open Label, Randomised Multicentre Phase III Study Of Irinotecan Hydrochloride (Campto (Registered)) And Cisplatin Versus Etoposide And Cisplatin In Chemotherapy Naive Patients With Extensive Disease - Small Cell Lung Cancer. NCT00143455.

id	Source (included studies)	Requested acknowledgments statements
		• An Open-label, Randomized, Phase 3 Clinical Trial of ABX-EGF Plus Best Supportive Care Versus Best Supportive Care in Subjects With Metastatic Colorectal Cancer. NCT00113763. • A Phase III, Randomised, Double-blind, Placebo-controlled Study to Assess the Efficacy and Safety of 10 mg ZD4054 (Zibotentan) in Combination With Docetaxel in Comparison With Docetaxel in Patients With Metastatic Hormone-resistant Prostate Cancer. NCT00617669. • A Multicenter, Randomized, Double-Blind, Phase 3 Study Of Sunitinib Plus Prednisone Versus Prednisone In Patients With Progressive Metastatic Castration-Resistant Prostate Cancer After Failure Of A Docetaxel-Based Chemotherapy Regimen. NCT00676650. All analyses and conclusions in this publication are the sole responsibility of the authors and do not necessarily reflect the opinions of the owners of the information, the clinical trial investigators, the NCTN, the NCORP, the National Cancer Institute, or Project Data Sphere. Neither the owners of the information, the clinical trial investigators, the NCTN, the NCORP, the National Cancer Institute, nor Project Data Sphere have contributed to, approved or are in any way responsible for the contents of this publication
8	https://biolincc.nhlbi.nih.gov/studies [98] (NCT00650091[99], NCT00000589[100], NCT00075829[101], NCT01982968[102], NCT01134783[103], NCT00004562[104])	This Manuscript was prepared using BMTCTN0102, CHOICES, PANTHER, OAT, TRAP, WRAP_IPF Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the BMTCTN0102, CHOICES, PANTHER, OAT, TRAP, WRAP_IPF or the NHLBI
9	https://nda.nih.gov/get/access-data.html [105] (NCT00012558[106], Registration not found[107], NCT01927276[108], NCT01944046[109],NCT00005013[110])	Data and/or research tools used in the preparation of this manuscript were obtained from the National Institute of Mental Health (NIMH) Data Archive (NDA). NDA is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in mental health. Dataset identifier(s): 3058, 2147, 2622, 2724, 2009, 2157. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or of the Submitters submitting original data to NDA.
10	https://vivli.org/ [111] (NCT01198756[112], NCT01573767[113], NCT01313676[114], NCT00400855[115], NCT01498822[116])	This publication is based on research using data from data contributors GSK and UCB that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.
11	https://beta.ukdataservice.ac.uk/datacatalogue/studies (reshare.ukdataservice.ac.uk) (https://www.ukdataservice.ac.uk/deposit-data) [117] (ISRCTN11288961[118], ISRCTN90749868[119] NCT01801410[120], Registration not required[121], ISRCTN24081411[122])	To Nottingham Clinical Trials Unit for facilitating sharing of data from the PRIDE study via ukdataservice.ac.uk
12	https://med.data.edu.au/find-data/ [123]	Repository no longer available
13	https://dcri.org/our-approach/data-sharing/soar-data [124] SOAR data: Available datasets: Duke cardiac catheterization datasets.	We acknowledge Michael Cohen-Wolkowiez, who is the principal investigator who conducted the original study from which the data were generated. Furthermore, we acknowledge the Eunice Kennedy Shriver National Institute of Child Health and Human Development Data and Specimen Hub for providing the Pharmacokinetics of Clindamycin and Trimethoprim-sulfamethoxazole in Infants and Children (PBPK) data that were used for this research.
14	https://journals.plos.org/plosone/search [125] (NCT02700490[126], TCTR20201005002[127], NCT02185196[128], ACTRN12616000538448[129], ISRCTN 71217488[130] NCT02747524[131])	No statement required
15	https://www.bmj.com/search/advanced [132] (NCT02068885[133], ,NCT01953549[134], ISRCTN11980540[135])	No statement required
16	https://dataverse.harvard.edu/ [136] (CTRI/2016/09/007240[137], PACTR201901905832601[138] NCT02148952[139], SLCTR/2019/015[140], ANZCTR12616001367437[141])	No statement required
17	https://arlg.org/studies-in-progress/ [142]	Not applicable as RCT data was not located in this repository
18	https://repository.niddk.nih.gov/studies/ [143]	Not applicable as RCT data was not located in this repository
Note: Note: These data-sharing acknowledgments are taken from Section 4.2 of this thesis and are included here for completeness. Their associated references can also be found in Section 4.2		

Contents

Index of tables and figures

Contribution to science

Peer review publications

Rodriguez A, Tuck C, Dozier MF, Lewis SC, Eldridge S, Jackson T, Murray A, Weir CJ. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Clinical Trials. 2022 Aug;19(4):452-63.

Rodriguez A, Lewis SC, Eldridge S, Jackson T, Weir CJ. A survey on UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trial datasets. Clinical Trials. 2024;0(0).

Submitted. Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications. Clinical Trials. 2024.

Invited presentations

Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. Using re-identification risk scores on publicly available anonymised clinical trial datasets. Oral presentation at the 7th International Clinical Trials Methodology Conference (Edinburgh, October 2024).

Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications. Oral presentation at the Statisticians in the Pharmaceutical Industry (PSI) 2024 conference (Amsterdam, June 2024).

Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Oral presentation at the Asthma UK Centre for Applied Research (AUKCAR) seminar (Edinburgh, 2024).

Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Oral presentation at the Centre Medical Informatics, The University of Edinburgh (Edinburgh, 2024).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Oral presentation at Women in data 2023 (Edinburgh, 2023).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Oral presentation at ECTU (Edinburgh, 2023)

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Oral presentation at 6th International Clinical Trials Methodology Conference (Harrogate, 2022).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Oral presentation at AUKCAR seminar (Edinburgh, 2021).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. Are anonymised databases truly anonymous? Lighting talk at AUKCAR seminar (Edinburgh, 2020).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. Are anonymised databases truly anonymous? Progress so far. Oral presentation at ECTU (Edinburgh, 2020).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? Patient Involvement review. Oral presentation for PPI AUKCAR members (Edinburgh, 2020).

Rodriguez A, Tuck C., Dozier MF, Mesa Eguiagaray I, Lewis SC, Eldridge S, Weir CJ. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Oral presentation at 5th International Clinical Trials Methodology Conference (Brighton, 2019).

Rodriguez A, Tuck C., Dozier MF, Mesa Eguiagaray I, Lewis SC, Eldridge S, Weir CJ. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Oral presentation at ECTU (Edinburgh, 2019).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An introduction. Oral presentation for 3-minute thesis at AUKCAR seminar (London, 2018).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An introduction. Oral presentation at AUKCAR seminar (Edinburgh, 2018).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An introduction. Oral presentation at Dealing with Data 2017 (Edinburgh, 2017).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An introduction. Oral presentation at AUKCAR research retreat (Edinburgh, 2017).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An introduction. Oral presentation at ECTU (Edinburgh, 2017).

Conference abstracts

Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. Using re-identification risk scores on publicly available anonymised clinical trial datasets. Submitted to 7th International Clinical Trials Methodology Conference (Edinburgh, April 2024).

Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets. Submitted to 7th International Clinical Trials Methodology Conference (Edinburgh, April 2024).

Rodriguez A, Lewis SC, Jackson T, Weir CJ, Eldridge S. A survey on UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets. Submitted to AUKCAR Annual Scientific Meeting (Reading, April 2024).

Rodriguez A, Lewis SC, Jackson T, Weir CJ, Eldridge S, UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets. Submitted to Statisticians in the Pharmaceutical Industry (PSI) 2024 conference (Amsterdam, November 2023).

Rodriguez A, Lewis SC, Jackson T, Weir CJ, Eldridge S, UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets. Submitted to Usher Institute Annual Lecture and Showcase 2023 (Edinburgh, September 2023).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Submitted to Women in data 2023 (Edinburgh, May 2023).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Submitted to AUKCAR Annual Scientific Meeting (Swansea, Feb 2023).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Submitted to Usher Institute Annual Lecture & Showcase (Edinburgh, October 2022).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Submitted to 6th International Clinical Trials Methodology Conference (Harrogate, October 2022).

Rodriguez A, Jackson T, Eldridge S, Lewis SC, Weir CJ. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Submitted to AUKCAR Annual Scientific Meeting (Leeds, June 2022).

Rodriguez A, Tuck C., Dozier MF, Mesa Eguiagaray I, Lewis SC, Eldridge S, Weir CJ. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Submitted to 5th International Clinical Trials Methodology Conference (Brighton, October 2019).

Rodriguez A, Tuck C., Dozier MF, Lewis SC, Eldridge S, Weir CJ. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Submitted to AUKCAR Annual Scientific Meeting (London, March 2019).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? Progress for year one. Submitted to AUKCAR – MRC Asthma UK Joint Meeting (London, September 2018).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An Introduction. Submitted to AUKCAR Annual Scientific Meeting (Bristol, January 2018).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An introduction. Submitted to Dealing with Data 2017 (Edinburgh, 2017).

Poster presentations

Rodriguez A, Williams LJ, Lewis SC, Sinclair P, Eldridge S, Jackson T, Weir CJ. UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets. At the 7th International Clinical Trials Methodology Conference (Edinburgh, October 2024).

Rodriguez A, Lewis SC, Jackson T, Weir CJ, Eldridge S. UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets. Presented at AUKCAR Annual Scientific Meeting 2024 (Reading, April 2024).

Rodriguez A, Lewis SC, Jackson T, Weir CJ, Eldridge S. UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets. Presented at Usher Institute Annual Lecture and Showcase 2023 (Edinburgh, September 2023).

Rodriguez A, Lewis SC, Jackson T, Weir CJ, Eldridge S. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? Presented at Usher Institute Annual Lecture & Showcase (Edinburgh, October 2022).

Rodriguez A, Tuck C., Dozier MF, Lewis SC, Eldridge S, Weir CJ. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Presented at AUKCAR Annual Scientific Meeting 2019. (London, March 2019).

Rodriguez A., Lewis SC, Weir CJ, Eldridge S. Are anonymised databases truly anonymous? An Introduction. Presented at AUKCAR Annual Scientific Meeting 2018. (Bristol, January 2018).

Abbreviations

Abbreviation	Term
ADaM	Analysis Data Model
ADF	Anonymisation Decision-Making Framework
APEC	Asia-Pacific Economic Cooperation
AUKCAR	Asthma UK Centre for Applied Research
CDISC	Clinical Data Interchange Standards Consortium
CMVM/UoE	The College of Medicine & Veterinary Medicine at the University of Edinburgh
COVID-19	Coronavirus disease, an infectious disease caused by the SARS-CoV-2 virus
CRAN	Comprehensive R Archive Network
CRF	Code of Federal Regulations
CROs	Contract Research Organisations
CSDR	Clinical Study Data Request websitewww.clinicalstudydatarequest.com
CSRs	Clinical Study Reports
CTUs	Clinical Trials Units
e.g.	Latin abbreviation for “exempli gratia” and means “for example”
ECTU	Edinburgh Clinical Trials Unit
EDPS	European Data Protection Supervisor
EEA	European Economic Area
EMA	European Medicines Agency
EMREC	Edinburgh Medical School Research Ethics Committee
ERO	Edinburgh Research Office
Et al.	Latin abbreviation for “et alia” and means “and others”
EU	European Union
FASDA	Fully Anonymous Secondary Data Analysis
FDA	Food and Drug Administration
GAPP	GaPP: a pilot randomised controlled trial of the efficacy of action of gabapentin for the management of chronic pelvic pain in women:
GCP	Good Clinical Practice
GDPR	General Data Protection Regulation
GMP	Good Manufacturing Practice
HHS	Health & Human Services
HIPAA	Health Insurance Portability and Accountability Act of 1996
HIV	Human immunodeficiency virus
HSS	The Department of Health & Human Services (US)
i.e.	Latin abbreviation for “id est” and means “in other words.”
ICH	The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICMJE	The International Committee of Medical Journal Editors
ICO	Information Commissioner’s Office
ICTMC	International Clinical Trials Methodology Conference
ICTRP	International Clinical Trials Registry Platform

Abbreviation	Term
IPD	Individual Participant data
IRBs	Institutional Review Boards
ISRCTN	International Standard Randomised Controlled Trial Number
JBI	Joanna Briggs Institute
MHRA	Medicines and Healthcare products Regulatory Agency
MRC	Medical Research Council
NHS	National Health Service
NIH	National Institutes of Health
NIHR	National Institute for Health Research
PhD	Doctorate of Philosophy is an abbreviation of the Latin phrase "philosophiae doctor"
PHI	Protected health information
PII	Personally identifiable information
PIPL	Personal Information Protection Law
PMDA	Pharmaceuticals and Medical Devices Agency
PPI	Patient and Public Involvement
RCT	Randomised controlled trial
SAGE Inc	A global academic publisher of books, journals, and a growing suite of library products and services.
SDTM	Study Data Tabulation Model
TOPPIC	The TOPPIC Trial: a randomised, double-blind parallel-group trial of mercaptopurine versus placebo to prevent recurrence of Crohn's disease following surgical resection in 240 patients
UK	United Kingdom
UKCRC	UK Clinical Research Collaboration
UoE	The University of Edinburgh
US	United States
USD	United States dollar
USS	Universities Superannuation Scheme
WHO	World Health Organization

Definitions

Due to the variety of terms and meanings used in literature on the subject of data anonymisation, this section explains the meaning of some key terms as they are used in this PhD thesis.

Concept	Definition
Anonymisation(Rodriguez, Tuck et al. 2022)	A data set would be considered anonymised if it has been de-identified and then subsequent data manipulation/steps have been taken to further protect the dataset, for example, if a privacy model has been applied (e.g., k-anonymity) or the link with the original non anonymised dataset has been destroyed and this action cannot be reversed.
Big Data(Google 2024)	Extremely large, diverse and complex datasets that cannot be handled by traditional data processing methods. It is characterized by high volume, velocity, and variety, often requiring advanced tools and techniques for analysis. The analysis of big data can reveal patterns, trends, and insights valuable for decision-making across various fields.
Brexit (Government of the Netherlands)	Name given to the United Kingdom's departure from the European Union. Brexit is the combination of the two English words: 'Britain' and 'exit'.
Clinical trials (NHS)	Research studies involving human participants to evaluate the safety and effectiveness of new medical treatments, drugs, devices, or interventions.
Clinical trial dataset	Clinical trial data organised into one or more tables of columns and rows with a relational database structure, where each table has a unique key (relation) identifying each row. Rows are also called records or tuples. Columns are also called attributes or variables. They will be assumed to be always two dimensional (rows by column structure) for this PhD.
Clinical Trial Units (UK Clinical Research Collaboration (UKCRC) 2023)	Specialised research facilities typically found within academic medical centres or hospitals. CTUs are dedicated to the planning, coordination, and conduct of clinical trials.
Controlled Access(Tudur Smith, Hopkins et al. 2015)	Datasets that can only be accessed if permission is granted by the data holders via their internal procedures.

Concept	Definition
De-identification	Removal of all personal health information and all other indirect identifiers which could lead to the identification of an individual. The most common de-identification methods are: 1. HIPAA (US Health Insurance Portability and Accountability Act of 1996) Safe harbour, in which 18 identifiers are removed from the datasets (U.S. Government 1996) (U.S. Department of Health & Human Services (HHS) 2012) 2. Hrynaszkiewicz et al. (Hrynaszkiewicz, Norton et al. 2010) proposed an enhanced removal of potential identifiers which are commonly present in clinical trials datasets.
Journalist scenario(El Emam 2013)	If the adversary sets out to identify any individual from the publicly available dataset just to prove that it can be done by using another dataset for "matching" with the publicly available dataset, then we are under journalist re-identification risk scores.
Matching dataset(El Emam 2013)	An independently obtained dataset with direct identifier, this dataset also needs to contain at least two matching variables to link it with the publicly available dataset.
Metadata	Data that describes characteristics and/or contents of anonymised/de-identified datasets.
Open Access(Tudur Smith, Hopkins et al. 2015)	Datasets that can be accessed without any or minimal restrictions imposed by the data holders.
Prosecutor scenario(El Emam 2013)	If the adversary knows that a target individual (for whom identifiers are known) is in the publicly available dataset (released anonymised and/or de-identified) we are under prosecutor re-identification risk scores. This scenario seeks to identify uniqueness in the records of the publicly available dataset.
Publicly available datasets	Data sets that are discoverable and available for sharing via open or controlled access, this data can be located on central repositories or with individual institutions/researchers.
Record	Records in a database or spreadsheet are also called rows, and it is a collection of fields, possibly of different data types. Usually each individual or an individual-visit combination is represented as a single record.

Concept	Definition
Re-identification risk score(El Emam 2013**)**	Estimated probability of any given individual being re-identified from an anonymised/de-identified dataset. The re-identification risk score depends on the variables available in the dataset, the number of observations in the dataset and on the strategy used to attack the dataset (prosecutor or journalist scenario).
Secondary research (Saunders, Lewis et al. 2009)	Also known as desk research, it involves analysing and synthesising existing data from previously published sources, rather than collecting new data. This type of research is used to conduct analysis that was not included in the original data collection proposal.
Variable	Also reference as attribute or column. A set of data values of a particular simple type, one value for each row of the database.

Outline of the thesis

This doctoral research was conducted part time over six years with the support of the University of Edinburgh (UoE) and the Asthma UK Centre for Applied Research (AUKCAR). I pursued this research concurrently with my role as a clinical trials statistician at the Edinburgh Clinical trials Unit (ECTU).

The genesis of this project stemmed from observations made by my supervisors, Steff Lewis and Chris Weir, who also serve as my line managers. They noted an increasing demand for sharing anonymised clinical trial data at ECTU, coupled with a dearth of knowledge in this particular domain. Subsequently, they initiated a competitive PhD funding opportunity through UoE and UKCAR to address this gap.

Drawing upon my experience at ECTU and my aspiration to advance my career as a clinical trial statistician through doctoral studies, I applied for the advertised PhD program and was successfully awarded in June 2017.

At the outset, the chosen topic 'Are anonymised datasets truly anonymous?' presented a broad scope with numerous avenues for exploration. As time progressed, we recognised its potential for controversy as well. My supervisors and I diligently worked to delineate a meaningful research focus aimed at addressing some of the gaps in our understanding of anonymisation and data sharing in clinical trials.

This PhD research encompasses three main complementary stages: a systematic scoping review, an analytical assessment of re-identification risks in publicly available clinical trial datasets, and an exploratory survey on UK researchers' opinions regarding data sharing. Each of these stages will be systematically outlined

in the subsequent chapters. To present this research comprehensively, the thesis has been organised into six chapters.

Chapter 1 outlines the finalised aims, objectives, scope and importance of this PhD.

Chapter 2 introduces this PhD by providing a comprehensive overview of clinical trials, including their regulatory framework, transparency requirements, and how their intersect with personal data regulations. It then delves further into the specific domain of sharing anonymised data for clinical trial datasets, setting the context for this research.

Chapter 3 details the methodology, execution, and findings of a systematic scoping review, consolidating guidelines and recommendations and providing an evidence base to guide subsequent stages of the project.

Chapter 4 reports on the development, execution, and outcomes of the assessment of re-identification risks associated with publicly available anonymised datasets.

Chapter 5 presents the methodology, implementation, and results of a cross-sectional survey aimed at gathering the views of UK researchers on de-identification, anonymisation, release methods, and re-identification risk estimation for clinical trial datasets.

Finally, in Chapter 6, I offer a comprehensive discussion summarising the conclusions and key findings from preceding chapters. Additionally, I explore avenues for future research and work in this field.

Chapter 1 Aims and objectives

1.1 Aims

This PhD aimed to develop recommendations for enhancing the privacy protection of clinical trials participants when sharing their anonymised datasets with external parties for further investigations. This involved understanding the requirements and characteristics necessary to declare a dataset anonymised in the context of clinical trials, estimating the re-identification risk of existing publicly available anonymised datasets, and investigating the views of researchers who undertake these anonymisation tasks.

1.2 Objectives of PhD

a. To describe the available anonymisation methods/techniques for clinical trials datasets.

b. To investigate whether individual participants could potentially be at risk of being re-identified from a range of datasets that have been anonymised and made available for sharing.

c. To identify factors that could increase the risk of re-identification of an anonymised clinical trial dataset.

d. To explore researchers' views and experiences regarding the sharing of clinical trial data.

e. To develop evidence-based recommendations on anonymisation techniques and data security for clinical trial datasets.

The work on this PhD adhered to the principles outlined by Erlich (Erlich 2013) in the editorial “Breaking Good: A Short Ethical Manifesto for the Privacy Researcher”: Therefore, I sought to:

1. "Increase the general knowledge" by having the findings of this PhD communicated via open access publication and conference proceedings despite limited resources
2. "Do no harm" as I did not attempt to identify or single out participants with this research or to bring any organisation into disrepute.
3. "If it is broken, try to fix it", findings and recommendations have been communicated via open access publication and conference proceedings.
4. "Don’t be over-confident", for this I had all results checked by a group of my peers.
5. "Not be afraid", here all methodology and results have been published under open access, we used all tools at our disposal to calculate the re-identification risk scores and we would welcome further requests for information.

1.3 Scope

This PhD research focused on the specific needs and implications of anonymising clinical trial datasets, setting itself apart from existing studies on big data (Bin, Jian et al. 2008, Potiguara Carvalho, Potiguara Carvalho et al. 2020) and routinely collected records(El Emam, Jabbouri et al. 2006) (Balas, Vernon et al. 2015), which extensively address anonymisation and de-identification. Although, audio-visual, imaging and genomic data are often included in clinical trial datasets, they were excluded from this study because they possess inherent characteristics (Aryanto, van Kernebeek et al. 2016) (Doel, Shakir et al. 2017) (Evans and Jarvik 2018) that do not respond well to anonymisation

techniques and privacy models designed for tabular data. The research also considered the limitations of the one-size-fits-all approach that data privacy regulators typically apply to data sharing and anonymisation. Additionally, I focused on clinical trial datasets, because this area aligns with my expertise, and I have easy access to the data. I believe this research can make a significant impact and could potentially be scaled to other areas that use personal data.

1.4 Importance of this PhD

Existing clinical trial data could be used to expand medical and scientific knowledge by exploring questions beyond the original study scope. It could facilitate individual participant data (IPD) meta-analysis, verify results, and investigate novel methodologies for data analysis. The importance and significant benefits of clinical trial data sharing have been well documented by many researchers (Gøtzsche 2011),(Packer 2016, Bertagnolli, Sartor et al. 2017),(Al-Shahi 2000, Pisani, Aaby et al. 2016).

The example given of data sharing between Copernicus and Kepler by Packer(Packer 2016) provides an excellent reason for why it is the right thing to do: Copernicus published his pivotal results demonstrating the existence of a heliocentric universe in 1543. Years later, Kepler would revisit Copernicus' data and discover discrepancies. These discrepancies provided Kepler with the evidence to support the theory that planets moved in an ellipse (instead of a circle) in 1604, an idea Kepler initially thought was too simple for earlier astronomers to have overlooked.

However, despite the clear benefits of data sharing, participant privacy protection must come first. One way to safeguard privacy is through anonymisation, but it must be done correctly. There are many documented cases in "big data" (large, routinely collected datasets) where individuals have been successfully re-identified from

released anonymised datasets (El Emam, Buckeridge et al. 2011, El Emam, Jonker et al. 2011, Sweeney 2013, Sweeney, Abu et al. 2013). These big data datasets are massive compared to clinical trials datasets. Therefore, we can infer that the re-identification risk for the former should be smaller than for the latter. This suggests that the threat of re-identification of anonymised individual patient data (IPD) from clinical trials, despite the lack of known cases, is a potential reality, particularly as traditional data sources (such as electoral rolls) and innovative ones (like social media) continue to evolve.

Not sharing data is no longer an option if we want to continue research, as funders (Cancer Research UK , National Institute for Health and Care Research (NIHR) 2019, Medical Research Council (MRC) 2023, National Institutes of Health (NIH) 2023) and publishers(Loder, Macdonald et al. 2024) (International Committee of Medical Journal Editors (ICMJE) 2024) are increasingly requesting the sharing of the data that back up the results from clinical trials. Hence, it is necessary to investigate ways to enhance the level of privacy protection provided to participants of clinical trials when anonymising datasets for sharing with the broader research community.

Chapter 2 Introduction

This chapter provides an overview of the context for this research.

2.1 Overview of Clinical trials

Clinical trials are fundamental to the advancement of medical science and the development of new treatments, medications, and medical procedures. They represent a systematic approach to evaluating the safety and efficacy of interventions in human subjects, aiming to provide reliable evidence for medical decision-making.

Understanding the definition and historical evolution of clinical trials provides insight into their significance and the rigorous processes involved.

2.1.1 History

The history of clinical trials dates back centuries, evolving from rudimentary observations to the rigorous scientific methodologies employed today. Early trials lacked standardised protocols and ethical considerations, often resulting in unreliable outcomes(Bhatt 2010).

The first recorded controlled clinical trial in modern time was done by the Scottish physician James Lind in 1747(Bhatt 2010). Lind conducted a trial to test treatments for scurvy among sailors, providing different groups with various supplements including citrus fruits. This study demonstrated the effectiveness of citrus fruits in preventing scurvy, laying the foundation for controlled clinical experimentation.

The 20th century saw advancements in the conduct of clinical trials, spurred by events like World War II, the development of the randomised controlled trial (RCT) (Bothwell, Greene et al. 2016) and tragedies such as the Tuskegee Syphilis Study (Cingi and Bayar Muluk 2016) and the thalidomide disaster (Bothwell, Greene et al. 2016). These

events were pivotal in the development of ethical frameworks such as the Nuremberg Code(U.S. Government Printing Office 1949), the Belmont report(Biomedical and Research 1978) and the Declaration of Helsinki(World Medical Association 2013) to guide human experimentation, protect trial participants, and continue to shape current clinical trial practices. Modern clinical trial regulation, overseen by agencies like US Food and Drug Administration (FDA), the European Medicines Agency (EMA), the UK Medicines and Healthcare products Regulatory Agency (MHRA) and the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan, emphasises methodological rigor, ethical oversight. and participant safety. Recent years have witnessed innovations in trial design, reporting and technology, aiming to enhance efficiency and data collection. Examples include the development of "patient reported outcome measures" (Black 2013) and the increasing use of cluster randomised trials (Moberg and Kramer 2015), to name just a few.

The history of clinical trials reflects a journey of evolving methodologies and ethical standards in medical research, from that first clinical trial conducted by Lind to fast-forward 275+ years, where clinical trials remain integral to medical advancements.

2.1.2 What are clinical trials

A clinical trial is a type of research study designed to evaluate the safety and efficacy of medical treatments, interventions, devices or procedures in humans (National Institute for Health and Care Research (NIHR) 2019). They "may involve patients, healthy people, or both" (NHS). These trials aim to gather scientific data to determine whether a new drug, therapy, medical device, or treatment approach is safe and effective for use in the general population; they could also aim to repurpose existing safe treatments. Typically, they compare the effect of a novel treatment against an existing treatment (called the control) or a placebo treatment (when no existing

treatment is available). Not every clinical trial will yield new and improved treatments. Some may find that the treatment under examination is ineffective or has side effects that are more severe than those of existing treatments. However, these data remain valuable for researchers, doctors, and ultimately, participants. (Cancer Research UK 2022)

Clinical trials typically follow a structured protocol or plan that outlines the objectives, methodology, participant eligibility criteria, treatment procedures, and outcome measures. They adhere to stringent regulatory oversight and strict ethical and scientific standards to protect the rights and welfare of participants, ensure research integrity, and uphold public trust. This includes obtaining informed consent from participants, ensuring confidentiality, and oversight by Institutional Review Boards (IRBs) or Ethics Committees. An aspect of particular importance is the use of randomised controlled trials (RCTs), a specific type of clinical trial where participants are randomly assigned to different treatment groups. This random allocation helps eliminate bias in the results, ensuring that any observed differences in the studied outcome are due to the treatment itself (MRC Clinical Trials Unit - UCL 2024).

2.1.3 Importance

Randomised controlled trials (RCTs) are crucial for advancing medical knowledge, improving patient care, and developing new treatments, ultimately, saving lives by identifying safe and effective treatments for various diseases and conditions (Fujimoto 2023). Due to their rigor, RCTs have proven to be the gold standard in medical research. This has positioned them at the top of the evidence-based medicine pyramid (Murad, Asi et al. 2016) (Figure 2-1).

Figure 2-1 The new evidence-based medicine pyramid.
Image source: Murad et al. 2016 (Murad, Asi et al. 2016)

Additionally, the global clinical trials market was valued at USD 48.68 billion in 2022 and is predicted to reach USD 83.55 billion by the year 2032 (Precedence Research Pvt Ltd 2024). Similarly, the number of trials, according to the World Health Organization (WHO) clinical trial registry, has been steadily increasing. By the end of 2022, it comprised a total of 577,531 studies, with 42,441 new studies added in 2022 alone (Figure 2-2).

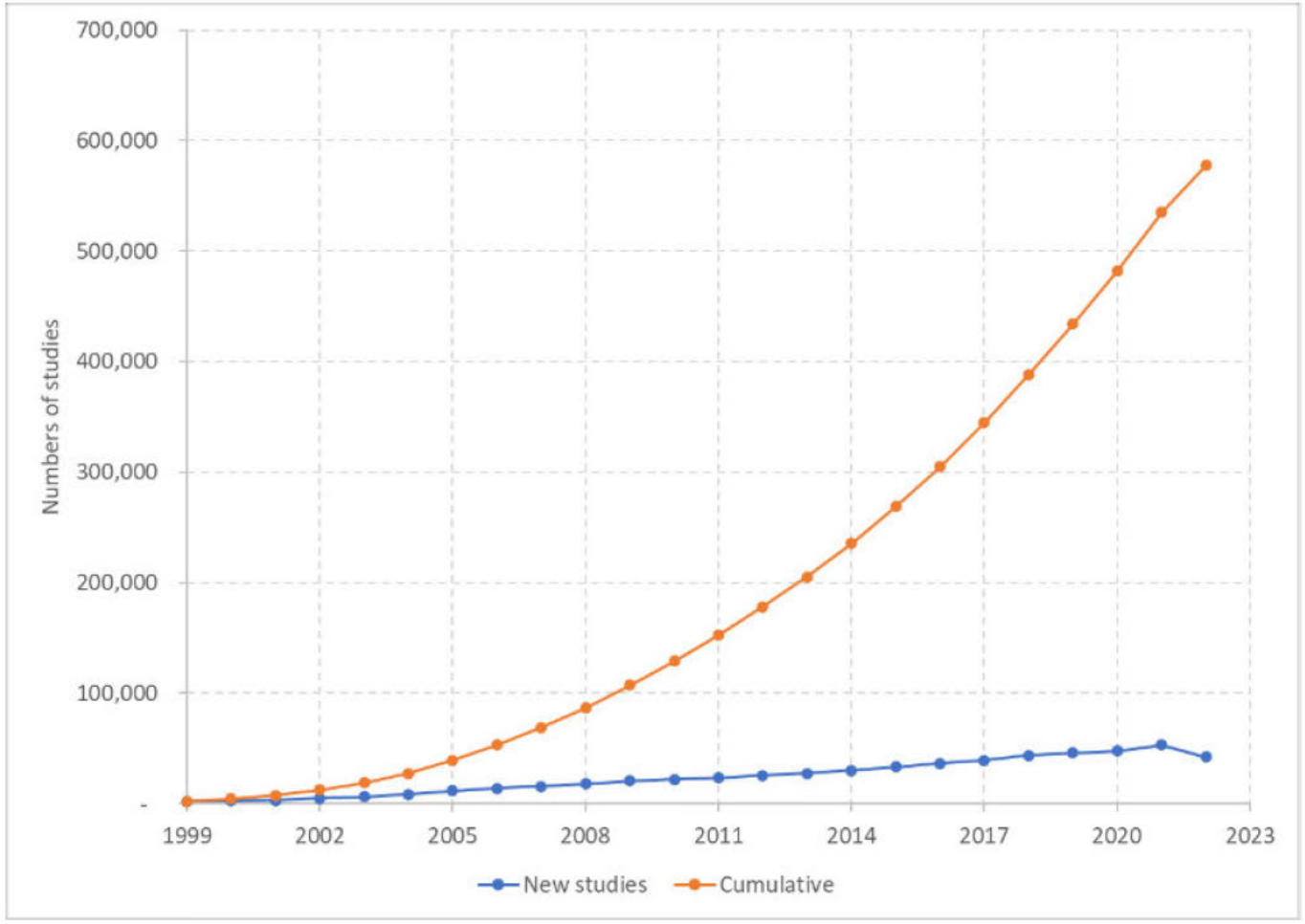


Figure 2-2 World Health Organization number of clinical trial registrations (1999-2022)

Image source: Original

Note: Created with data from the World Health Organisation WHO 2023 (World health Organization (WHO) 2023)

The largest share of the clinical trials market is in the North America region at 51.7%, followed by Europe and the Asia Pacific regions with 26.5% and 18.3%, respectively. (Precedence Research Pvt Ltd 2024). Various factors are driving the expansion of the global clinical trials market, including the rising prevalence of chronic disorders, the increasing number of clinical trials in developing regions, the growing availability of biologics, the heightened demand for advanced treatments like personalised medicines, the occurrence of viral disease outbreaks, the surge in global cancer cases, the aging population, and the escalating expenditure on research and development. (Precedence Research Pvt Ltd 2024).

2.1.4 Phases of Clinical Trials

Clinical trials for medicinal products are conducted in phases, starting from initial testing in a small group of people (phases I-II) and progressing to larger scale studies to assess the effectiveness and safety of the intervention (phases III-IV) (Evans 2010). Each phase serves a specific purpose (Figure 2-3):

- Phase 1 trials involve a small number of healthy volunteers and in certain instances patients (e.g., cancer phase I trials) and focus primarily on assessing the safety, dosage, and potential side effects of the intervention.
- Phase 2 trials enrol a larger group of participants, often patients with the targeted disease or condition, to further evaluate safety and begin assessing efficacy.
- Phase 3 trials involve a larger and more diverse population of patients and are designed to provide more comprehensive data on safety and efficacy. These trials compare the new treatment with existing standard treatments or a placebo (inactive substance) to determine its effectiveness and potential benefits.

- Phase 4 trials, also known as post-marketing surveillance trials, phase 4 trials occur after a treatment has been approved for use by regulatory agencies. They continue to monitor the treatment's safety and effectiveness in a real-world setting and may identify rare side effects or long-term effects that were not evident in earlier phases.

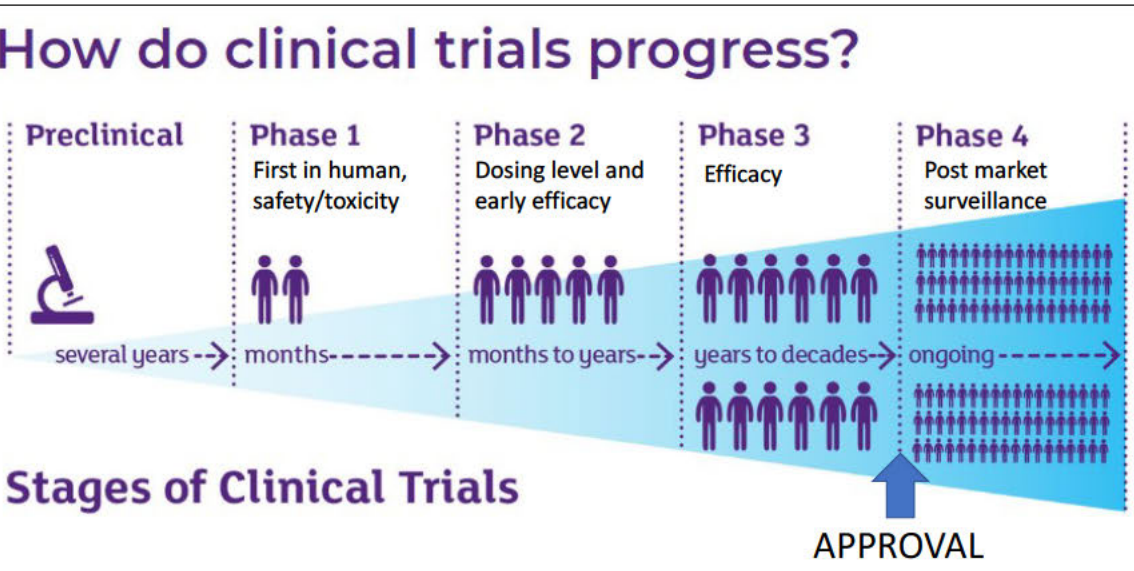


Figure 2-3 Phases of clinical trials

Image source: © State of New South Wales (Cancer Institute NSW). For current information go to www.cancer.nsw.gov.au (Cancer Institute New South Wales (NSW) 2022)

Permission given by the author to reproduce this content

Typically, a new treatment must undergo several phase 3 clinical trials before researcher and regulators are confident enough to adopt it as the new standard treatment. While a positive trial outcome may occur by chance or due to a poorly designed trial, the likelihood diminishes when multiple trials produce consistent results. (Cancer Research UK 2022). It is important to note that some clinical trials do not follow the defined phases described, such as trials involving behavioural interventions or medical devices; these are classified as phase "not applicable" (Dal-Ré, Banzi et al. 2023).

2.1.5 How are trials executed?

The lifecycle of a clinical trial encompasses the entire process from study conception to completion and beyond. This includes, for example, protocol development, participant recruitment, data collection and analysis, regulatory submissions, and dissemination of results. Each stage of the lifecycle requires meticulous planning, execution, and monitoring to ensure scientific integrity, participant safety, and regulatory compliance (Figure 2-4).

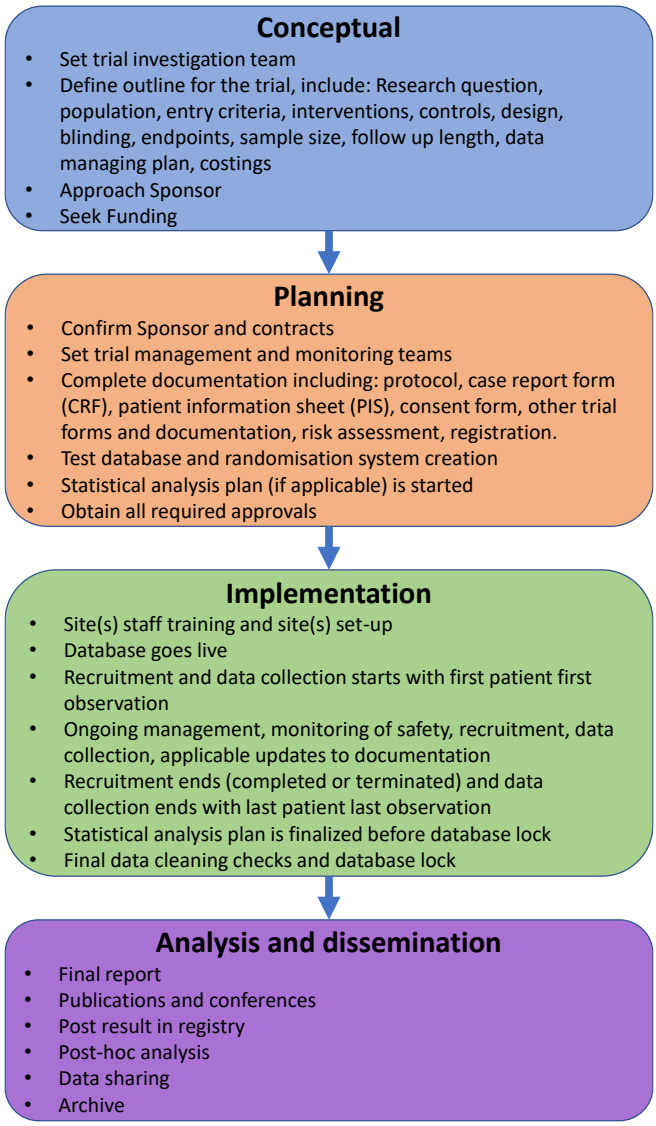


Figure 2-4 Lifecycle of a clinical trial

Image source: Adapted from Bagley et al. 2016 (Bagley, Short et al. 2016), with information from "Fundamentals of clinical trial design" by Evans 2010 (Evans 2010)

Clinical trials begin as an idea and conclude with results, yielding deliverables such as clinical study reports (CSRs), publications and data. These deliverables could subsequently be used for regulatory approval, the development or updating of clinical practice guidelines, the updating of drug uses and their side effects, and to inform healthcare decision-making, just to name a few.

2.1.6 Who Executes Clinical trials

Clinical trials are typically conducted by a variety of entities including (but not limited to):

- Pharmaceutical companies and biotech firms: These organisations often initiate and sponsor clinical trials to test the safety and efficacy of new drugs, therapies, or medical devices.
- Academic institutions: Universities and medical schools often conduct clinical trials, sometimes in collaboration with pharmaceutical companies, government agencies, non-profit organisations, hospitals and medical centres. These trials may be led by researchers or clinicians with expertise in the relevant field.
- Government agencies: Government agencies such as the National Institutes of Health (NIH) in the United States, the European Medicines Agency (EMA) in Europe, the National Institute for Health Research (NIHR) in the UK or similar organisations in other countries may fund and conduct clinical trials, particularly for investigating public health concerns or evaluating treatments for rare diseases.
- Contract research organisations (CROs): These are independent organisations hired by pharmaceutical companies or other sponsors to manage various aspects of clinical trials, including protocol design, participant recruitment, data management, and regulatory compliance.

- Hospitals and medical centres: Clinical trials are often conducted within hospitals and medical centres, where clinicians and researchers have direct access to patients and resources for conducting studies.
- Patient advocacy groups: Sometimes, patient advocacy groups initiate or sponsor clinical trials to address specific medical conditions or to evaluate treatments from a patient perspective.
- Non-profit organisations: Non-profit organisations may conduct clinical trials as part of their mission to improve healthcare outcomes or to address specific medical needs in underserved populations.

Clinical trials involve collaboration among various stakeholders, including healthcare professionals, researchers, regulatory authorities, and participants, to ensure the ethical conduct and scientific validity of the studies. Occasionally, these groups formalise their collaboration by establishing clinical trial units (CTUs).

2.1.7 What are clinical trial units?

Clinical trial units (CTUs) are specialised organisations dedicated to conducting clinical trials. These units are often found within academic medical centres, hospitals, research institutions, or private organisations and are staffed by multidisciplinary teams of researchers, clinicians, coordinators, and support staff.

Key features of clinical trial units include:

- Trial Management: CTUs oversee the planning, implementation, and coordination of clinical trials from inception to completion. They develop protocols, recruit participants, manage study logistics, and ensure compliance with regulatory requirements and ethical standards.
- Clinical Expertise: CTUs often have access to a diverse range of clinical expertise, including physicians, nurses, pharmacists, and other healthcare

professionals. This expertise ensures that clinical trials are conducted safely and effectively, with appropriate medical oversight and patient care.

- Methodology Expertise: CTUs also have experts in clinical trial methodology, such as statisticians and health economists, who play crucial roles in the design, analysis, and interpretation of data. Their contributions ensure that study results are scientifically valid and ethically sound. Both statisticians and health economists are integral to ensuring that clinical trials not only produce scientifically rigorous results but also provide practical insights into the value and impact of new medical interventions on both patients and healthcare systems.
- Research Infrastructure: CTUs provide access research infrastructure and facilities necessary for conducting clinical trials. This may include clinical research units, laboratory facilities, imaging centres, and data management systems to support data collection, analysis, and storage.
- Quality Assurance: CTUs prioritise quality assurance and adhere to Good Clinical Practice (GCP) guidelines and other regulatory standards to ensure the integrity, reliability, and ethical conduct of clinical trials. They may implement quality control measures, conduct site monitoring visits, and participate in audit and accreditation processes.
- Training and Education: CTUs often offer training and educational programs for research staff, investigators, and study participants to enhance knowledge and skills in clinical research methods, regulatory compliance, and ethical principles.
- Collaboration and Networking: CTUs collaborate with sponsors, industry partners, regulatory agencies, and other research institutions to facilitate the design, conduct, and dissemination of clinical trials. They may participate in

multicentre studies, consortia, and collaborative research networks to leverage resources and expertise.

CTUs play a critical role in advancing medical research and translating scientific discoveries into clinical practice. By providing infrastructure, expertise, and support for clinical trials, CTUs contribute to the development of new treatments, therapies, and interventions that improve patient care and outcomes.

In the UK, CTUs and the UK Clinical Research Collaboration (UKCRC) (UKCRC 2024) have a synergistic relationship aimed at enhancing the quality and efficiency of clinical research in the UK. The UKCRC accredits CTUs that meet rigorous standards of clinical trial management and conduct. This accreditation process ensures that CTUs have the necessary expertise, resources, and infrastructure to design, conduct, and analyse high-quality clinical trials. By fostering a network of accredited CTUs, the UKCRC promotes best practices, collaboration, and the sharing of knowledge and resources among units, thereby supporting the overall goals of improving clinical research and patient outcomes(UKCRC 2024).

2.2 Regulation of Clinical trials

2.2.1 The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH)

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) is a global organisation that brings together regulatory authorities and pharmaceutical industry representatives from around the world to develop and promote harmonised guidelines for the pharmaceutical industry. The ICH was established in 1990 as a joint initiative between regulatory authorities and industry associations from Europe (which included the UK), Japan,

and the United States. Since then, it has expanded to include regulatory agencies and industry groups from other regions, such as UK, Canada and Switzerland. The primary goal of the ICH is to facilitate the development and registration of safe, effective, and high-quality pharmaceutical products while reducing duplication of efforts and unnecessary regulatory barriers. (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2024) The key objectives of the ICH are:

- Harmonisation of Regulatory Requirements: The ICH develops guidelines and standards that aim to harmonise regulatory requirements for the development, registration, and post-approval oversight of pharmaceutical products. Harmonisation helps streamline the regulatory process, reduce delays in drug approvals, and facilitate global drug development and marketing.
- Guideline Development: The ICH develops guidelines covering wider aspects of pharmaceutical development, including quality, safety, efficacy, and multidisciplinary topics. These guidelines are developed through a consensus-driven process involving regulatory authorities and industry experts from member countries. Some of the notable guidelines developed by the ICH cover topics such as Good Clinical Practice (GCP), Good Manufacturing Practice (GMP), pharmacovigilance, clinical safety evaluation, and quality risk management, among others.
- Implementation Support: The ICH provides support and guidance to regulatory authorities and industry stakeholders to facilitate the implementation of its guidelines. This includes training programs, workshops, and other educational initiatives aimed at promoting understanding and adoption of ICH standards.

- Periodic Review and Revision: The ICH periodically reviews and updates its guidelines to ensure they remain current and reflect advancements in science, technology, and regulatory practices. This iterative process helps maintain the relevance and effectiveness of ICH standards over time.

The ICH plays a critical role in promoting global regulatory harmonisation and facilitating the development and registration of pharmaceutical products worldwide, ultimately benefiting participants by ensuring access to safe, effective, and high-quality medicines.

2.2.2 ICH guidelines for clinical trials

The core guidelines for clinical trials are outlined in ICH E8 (R1) (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2021) and ICH E6 (R2) (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2016). ICH E8 (R1) provides comprehensive principles guiding the design, conduct, recording, and reporting of clinical trials, emphasising ethical considerations, scientific integrity, and regulatory compliance across all phases of clinical development. This guideline covers various aspects of trial design, including protocol development, the selection of study populations, endpoints, and statistical considerations. Serving as a foundational document, ICH E8 (R1) assists sponsors, investigators, and regulatory authorities in conducting and evaluating clinical trials effectively.

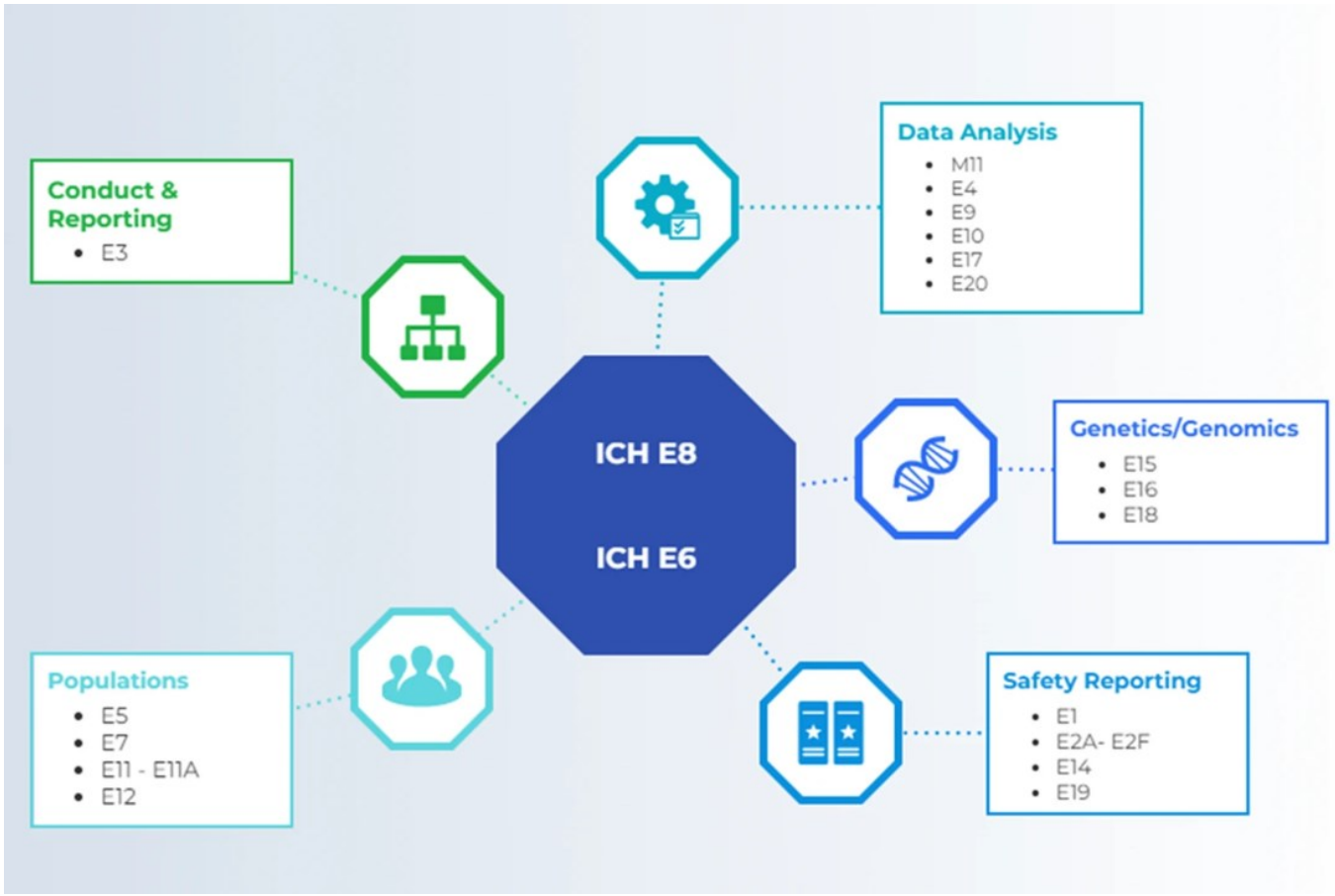
Meanwhile, ICH E6 (R2), also known as "Good Clinical Practice: Consolidated Guideline" or ICH GCP, establishes unified standards for the design, conduct, monitoring, auditing, recording, analysis, and reporting of clinical trials involving human subjects. Its primary focus is ensuring the protection of trial participants'

rights, safety, and well-being, as well as maintaining the credibility and accuracy of the generated data. ICH E6 (R2) plays a pivotal role in upholding the quality and integrity of clinical trial data by outlining the 13 key principles for the conduct of clinical trials (Table 2-1).

1	Ethics	“Clinical trials should be conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki, and that are consistent with GCP and the applicable regulatory requirement(s).”
2	Trial risk vs trial benefit	“Before a trial is initiated, foreseeable risks and inconveniences should be weighed against the anticipated benefit for the individual trial subject and society. A trial should be initiated and continued only if the anticipated benefits justify the risks.”
3	Trial participants	“The rights, safety, and well-being of the trial subjects are the most important considerations and should prevail over interests of science and society.”
4	Information on the Medicinal Product	“The available non-clinical and clinical information on an Investigational Product should be adequate to support the proposed clinical trial.”
5	Good quality trials	“Clinical trials should be scientifically sound, and described in a clear, detailed protocol.”
6	Compliance with the study protocol	“A trial should be conducted in compliance with the protocol that has received prior institutional review board (IRB)/independent ethics committee (IEC) approval/favourable opinion.”
7	Medical decisions	“The medical care given to, and medical decisions made on behalf of, subjects should always be the responsibility of a qualified physician or, when appropriate, of a qualified dentist.”
8	Trial staff	“Each individual involved in conducting a trial should be qualified by education, training, and experience to perform his or her respective task(s).”
9	Informed consent	“Freely given informed consent should be obtained from every subject prior to clinical trial participation.”
10	Clinical trial data	“All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification. ADDENDUM: This principle applies to all records referenced in this guideline, irrespective of the type of media used.”

11	Confidentiality	“The confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).”
12	Good Manufacturing Practice	“Investigational products should be manufactured, handled, and stored in accordance with applicable good manufacturing practice (GMP). They should be used in accordance with the approved protocol.”
13	Quality assurance	“Systems with procedures that assure the quality of every aspect of the trial should be implemented. ADDENDUM: Aspects of the trial that are essential to ensure human subject protection and reliability of trial results should be the focus of such systems.”
Table 2-1 The 13 principles of ICH GCP		
Data Source: ICH (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2016)		

These principles ultimately facilitate the development and registration of safe and effective pharmaceutical products, interventions, medical devices and therapies. Moreover, numerous other ICH guidelines complement E8 and E6, synergistically supporting the effective execution of clinical trials, as depicted in Figure 2-5



Where:

E6 Good Clinical Practice
E8 General considerations for clinical studies

Design Analysis
M11 Clinical electronic Structured Harmonised Protocol (CeSHarP)
E4 Dose–Response Studies
E9 Statistical Principles for Clinical Trials
E10 Choice of Control Groups in Clinical Trials
E17 Multi-Regional Clinical Trials
E20 Adaptive Clinical Trials

Conduct & Reporting
E3 Clinical Study Reports

Safety Reporting
E1 Clinical Safety for Drugs used in Long-Term Treatments
E2A–E2F Pharmacovigilance
E14 Clinical Evaluation of QT
E19 Safety Data Collection

Populations
E5 Ethnic Factors
E7 Clinical Trials Geriatric Population
E11–E11A Clinical Trials in Paediatric Population
E12 Clinical Evaluation by Therapeutic Category

Genetics/Genomics
E15 Definitions in Pharmacogenetics/Pharmacogenomics
E16 Qualification of Genomic Biomarkers
E18 Genomic Sampling

Figure 2-5 ICH Guidelines connected by E8 and E6

Image source: Whitehead et al. 2024 (Whitehead, Suprin et al. 2024)

ICH E9 Statistical Principles for Clinical Trials deserves special mention because it provides internationally recognised guidelines on the statistical methodology necessary to ensure the validity and reliability of clinical trial results. These principles help in designing, conducting, analysing, and reporting clinical trials, guaranteeing that the data collected is robust and the conclusions drawn are scientifically sound. Adhering to ICH E9 helps in maintaining the integrity of the research and ensures that the findings are credible and can be used to make informed decisions in clinical practice and policy.

2.2.3 Implementation of ICH

Regulatory agencies in different countries or regions, such as the US Food and Drug Administration (FDA), the European Medicines Agency (EMA), the UK Medicines and Healthcare products Regulatory Agency (MHRA) and the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan (Brody 2016), just to mention few, are responsible for overseeing the implementation of ICH guidelines within their respective jurisdictions. They may adopt, adapt, or harmonise ICH guidelines into their regulatory frameworks to ensure the safety, efficacy, and quality of pharmaceutical products. But in general, the implementation of ICH guidelines typically involves multiple stakeholders and is a collaborative effort involving regulatory authorities, pharmaceutical industry, healthcare professionals, and other relevant entities to ensure the safety, efficacy, and quality of pharmaceutical products and clinical research.

2.3 Clinical trial data transparency

Clinical trial data transparency refers to the practice of making clinical trial data and results accessible to various stakeholders, including researchers, healthcare professionals, regulatory agencies, and the public. It involves sharing detailed information about the design, conduct, findings, and outcomes of clinical trials in a timely and comprehensive manner. (Bruckner 2017)

The principles outlined in ICH E6(R2) currently support the goals of clinical data transparency by emphasising the importance of data integrity, reliability, and accuracy. Adhering to ICH E6(R2) guidelines helps ensure that clinical trial data is of high quality and can be effectively communicated to stakeholders. However, ICH E6(R2) does not explicitly recommend clinical trial data transparency, and as a result, commitments to transparency vary widely. For example, commercial sensitivity of data can limit willingness to share data due to concerns about protecting proprietary information, maintaining competitive advantage, and safeguarding intellectual property rights. This could particularly impact pharmaceutical companies and other industry sponsors who invest significant resources in research and development and may view their results and data as confidential assets (Ross, Tse et al. 2012, Hartung, Zarin et al. 2014, Goldacre, Lane et al. 2017, Rowhani-Farid, Grewal et al. 2023).

The explanatory note of ICH E6(R2) (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2021) (and its next version, ICH E6(R3) currently under public consultation (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2023)) are covering this gap, by addressing the transparency of clinical trials with the statement "The transparency of clinical trials in drug development includes registration on publicly accessible and recognised databases, and the public posting of clinical trial

results" (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2021) (The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) 2023). It is important to note that while ICH E6 focus is on generating reliable results, data sharing is still not explicitly included in the aforementioned statement.

2.3.1 Key aspects of clinical trial data transparency

Clinical trial researchers have generated various guidelines to foster data transparency. For instance, Goldacre et al. 2017 (Goldacre, Lane et al. 2017) proposed that the following information about a clinical trials should be shared: registration, methods and summary results, clinical study reports (CSRs) and individual participant data (IPD), this is also echoed on the report by Bruckner 2017 (Bruckner 2017), who recommended five fundamental items (Figure 2-6) for clinical trial transparency to be required by policy makers.

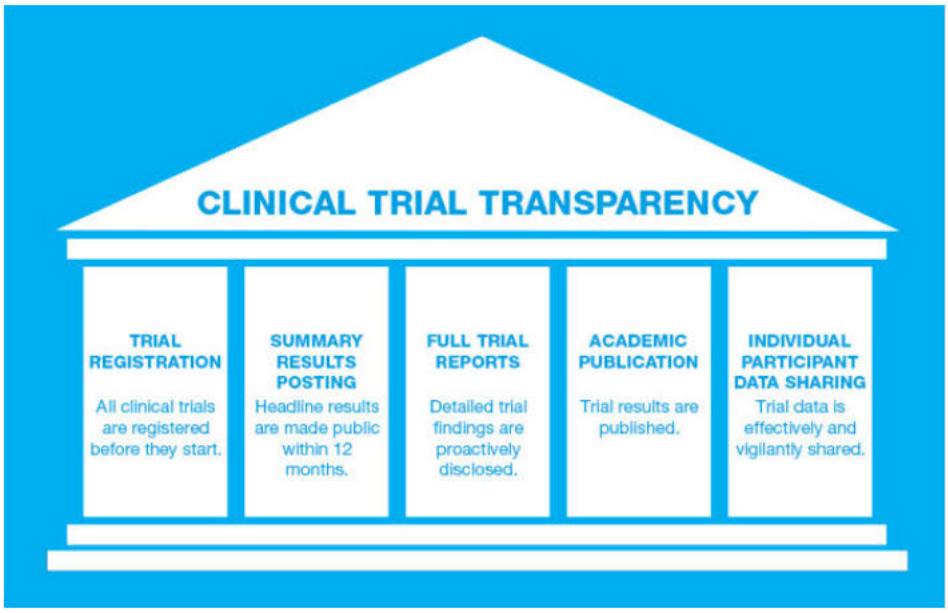


Figure 2-6 The five pillars of clinical trial transparency

Image source: Bruckner 2017 (Bruckner 2017)

2.3.1.1 Registration of trials

Clinical trial registration is the process of publicly documenting the details of a clinical trial before enrolling participants. It involves submitting comprehensive information about the trial to a publicly accessible registry. Clinical trial registration serves several important objectives. Firstly, it promotes transparency by making information about clinical trials publicly available, allowing stakeholders such as researchers, healthcare providers, participants, and policymakers to access essential details about ongoing and completed trials, such as the trial's purpose, design, eligibility criteria, and contact details. Secondly, registration helps prevent duplication of research efforts by enabling researchers to identify ongoing trials and assess whether similar studies are already underway. Thirdly, registration enhances ethical standards in clinical research by ensuring that study protocols are publicly documented before enrolling participants, thereby reducing the risk of selective reporting and outcome bias. Fourthly, registered trials contribute to the availability of comprehensive data for systematic reviews and meta-analyses, enabling researchers to synthesise evidence and draw more robust conclusions about the effectiveness and safety of interventions. Finally, registration fosters public trust and confidence in the research enterprise by demonstrating a commitment to transparency and accountability in clinical research practices.

While ICH E6 doesn't specifically mandate trial registration, it emphasises principles such as transparency, data integrity, and compliance with regulatory requirements, which align with the objectives of trial registration. In many countries, regulatory authorities or professional organisations mandate clinical trial registration as a requirement for conducting research involving human participants and as part of their oversight process to ensure that clinical trials are conducted ethically and transparently. Electronic registries such as the WHO International Clinical Trials

Registry Platform (ICTRP) (World health Organization (WHO)), the US National Library of Medicines ClinicalTrials.gov (NIH U.S. National Library of Medicine 2024) and BioMed Central International Standard Randomised Controlled Trial Number (ISRCTN) (BioMed Central Ltd) registry serve as central repositories for registered clinical trials, facilitating global access to trial information.

Key components of clinical trial registration typically include:

- Trial Protocol: Detailed description of the study objectives, design, methodology, and planned analyses.
- Participant Eligibility Criteria: Criteria defining who can participate in the trial, such as age, gender, medical condition, and other relevant factors.
- Interventions: Description of the experimental treatments or interventions being tested, including dosage, duration, and mode of administration.
- Comparators: A description of the standard treatment/intervention or placebo/sham (when no standard exists) used for comparison.
- Outcome Measures: Specification of the primary and secondary outcomes that will be evaluated to assess the effectiveness and safety of the interventions.
- Study Locations: Information about the sites where the trial will be conducted, including contact details for principal investigators and study coordinators.
- Funding and Sponsorship: Disclosure of sources of funding and the organisation or entity responsible for overseeing the trial.

2.3.1.2 Dissemination of Results

After the completion of a clinical trial, researchers are expected to publish the trial's results, which includes, academic publications, full trial reports (such as clinical study

reports (CSRs)) and summary results. Publication ensures that the findings of the trial are disseminated and are available for scrutiny, validation, and incorporation into medical knowledge and practice. Also, there is recognition of the importance of involving patients and the public in clinical trial data transparency efforts. Providing accessible and understandable information about clinical trials and their results empowers patients to make informed decisions about their healthcare and enhances public trust in the research enterprise (Palmer 2024).

2.3.1.3 Individual participant data (IPD)

In addition to publishing results, there is a growing emphasis on sharing individual participant data (IPD) for further analysis and secondary research (i.e., research not included in the original proposal for data collection). Data sharing promotes transparency, reproducibility, and collaboration in scientific research, allowing independent validation of findings and maximising the utility of collected data. Existing clinical trials research data could be used to expand medical and scientific knowledge, thereby improving 'patient outcomes' and 'public trust in regulatory agencies and medicines' (Modi, Kichenadasse et al. 2023) (Figure 2-7).

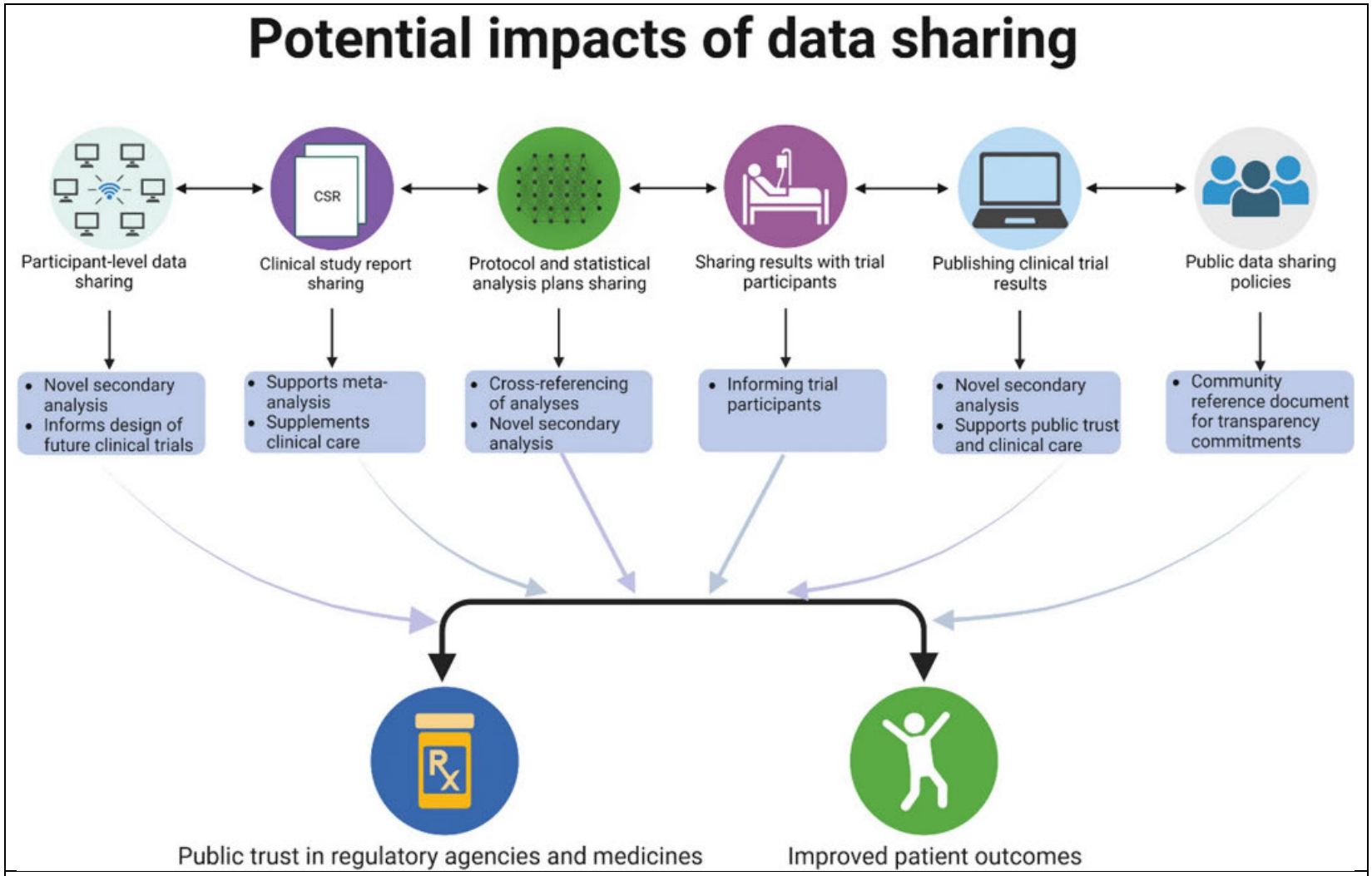


Figure 2-7 Potential impacts of data sharing

Image source: Modi et al. 2023 (Modi, Kichenadasse et al. 2023)

Similar to registration processes, funders, professional organisations, and regulatory agencies have been establishing guidelines and policies to promote transparency regarding the dissemination of results and the sharing of Individual Participant Data (IPD) (Zemła-Pacud and Lenarczyk 2023) (Kaplan, Koong et al. 2023) (Gamertsfelder, Figueroa et al. 2023) (Califf 2023). In many cases, regulatory agencies may require sponsors to submit detailed study protocols, summary results, and individual participant-level data for regulatory review and approval. Additionally, sponsors and regulatory agencies often aim to facilitate sharing with a broader audience, though this remains a developing area. Before delving into the topic of requests for sharing of clinical trials results and IPDs, it is important to note that harmonising the data is desirable to maximise the objective of achieving clinical trial data transparency (Mann, Pedersen et al. 2023).

2.3.2 Harmonisation of clinical trial data

The harmonisation of clinical trial data refers to the process of standardising and aligning data collection, management, and reporting practices across different clinical trials. This effort aims to increase transparency in clinical research by making trial data more consistent, comparable, and accessible. Currently, it is led by the Clinical Data Interchange Standards Consortium (CDISC) (Clinical Data Interchange Standards Consortium), a global non-profit organisation that develops and maintains data standards for clinical research.

Founded in 1997, CDISC aims to develop global, platform-independent data standards that ensure the interoperability of clinical research data and streamline regulatory review processes. By standardising the collection, exchange, and submission of clinical trial data, CDISC improves efficiency, quality, and collaboration across the clinical research process. Their standards cover various aspects of

clinical research, including study design, data capture, terminology, data management and analysis, ensuring consistent data reporting regardless of the technology used. This facilitates seamless data exchange and collaboration among all stakeholders involved in clinical research, reduces errors, inconsistencies, and redundancies, and enhances the reliability and robustness of clinical trial findings, ultimately accelerating drug development.

Key standards developed and maintained by CDISC include:

- Study Data Tabulation Model (SDTM) (Clinical Data Interchange Standards Consortium): A standard format for organising and presenting clinical trial data in a tabular format for submission to regulatory agencies.
- Analysis Data Model (ADaM) (Clinical Data Interchange Standards Consortium): A standard format for analysing and reporting clinical trial data, enabling consistent data analysis and interpretation.
- Controlled Terminology (Clinical Data Interchange Standards Consortium): A standardised set of terms and definitions used to describe clinical trial data elements and concepts.

CDISC standards are widely recognised and endorsed by regulatory agencies worldwide, including the US Food and Drug Administration (FDA), the European Medicines Agency (EMA), the UK Medicines and Healthcare products Regulatory Agency (MHRA), and the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan, who often require compliance with these standards for the submission of regulatory applications and marketing authorisation of new drugs and medical devices.

CDISC plays a vital role in promoting data standardisation and interoperability in clinical research, contributing to greater efficiency, transparency, and reliability in the

development and evaluation of new medical treatments. Specifically, by adhering to CDISC standards, researchers and organisations can enhance transparency in clinical trial data, ensuring that it is collected, stored, and reported consistently and in a format that facilitates sharing and analysis.

2.3.3 Request for the sharing of results and IPD

Humans are currently enjoying a time where information is at their fingertips, and its value can be maximised due to the great advancements in computational power, the internet, and the burgeoning field of Artificial Intelligence. Combined with the increasing demand for data transparency, this has created optimal conditions for data sharing. Therefore, regulators have taken steps to foster the sharing of results and Individual Participant Data (IPD) from clinical trials.

For example, the EMA released the "European Medicines Agency Policy on Publication of Clinical Data for Medicinal Products for Human Use - POLICY/0070" (European Medicines Agency (EMA) 2019) to provide a legal basis for the release of results. It is divided into two phases: the first phase concentrates on the publication of Clinical Study Reports (CSRs), while the second phase focuses on reviewing several aspects to implement the sharing of Individual Participant Data (IPD).

Similarly, the FDA uses 42 CFR Part 11, "Clinical Trials Registration and Results Information Submission" (National Institute of Health (NIH) - Department of Health and Human Services (HHS) 2016) to monitor and enforce the registration and publication of summary results. However, compliance with the rule is low (Anderson, Chiswell et al. 2015).

Regarding the sharing of IPDs, 42 CFR Part 11 only provides guidance and recommendations without legally enforceable responsibilities. Both Policy 0070 and 42 CFR Part 11 are still facing barriers for widespread implementation (Zemła-Pacud and Lenarczyk 2023) (Ferran and Nevitt 2019) (HMA Permanent Secretariat 2021) (Chaturvedi,

Mehrotra et al. 2019) (Dal-Ré and Mahillo-Fernández 2023) (Goldacre, DeVito et al. 2018) (Pisternick-Ruf, Marquart et al. 2018) (Holtedahl 2020) (Hanson 2018) (Vorland, Brown et al. 2024)

Meanwhile, The MHRA published the results of a consultation on legislative proposals for clinical trials (Medicines and Healthcare products Regulatory Agency (MHRA) 2023) in 2023. In this report, the MHRA proposed addressing data transparency by including requirements such as the "Requirement to register a trial", the "Requirement to publish a summary of results within 12 months of the end of the trial unless a deferral has been agreed" and the "Requirement to share trial findings with participants in a suitable format" through the drafting of new legislation. However, Law et al. (Law, Couturier et al. 2023) observed that the topic of sharing Individual Participant Data (IPD) was omitted. Therefore, there are gaps in transparency guidance from regulators regarding data sharing; as a result, it is important to stay up to date with their rapidly evolving requirements and recommendations (PHUSE 2020). (Eichler and Rasi 2020)

In the same way, there has also been an increase in the number of requirements from funders and publishers, for clinical researchers to share their research data with others, within the existing legal framework, once the primary analysis has been completed. For example, new grant applications with funding from Cancer Research UK(Cancer Research UK) and the Medical Research Council(Medical Research Council (MRC) 2023) must contain a concrete data-sharing plan. The funding bodies(Digital Curation Centre) have also released guidance to create data sharing plans and how to estimate the cost related with data sharing activities and, at the start of 2018, the Medical Research Council, along with the Wellcome Trust, Cancer Research UK, and the Bill and Melinda Gates Foundation, announced their commitment to covering all expenses associated with sharing academic clinical trial data through the Clinical Study Data Request (CSDR) website(Digital Curation Centre)

Regarding publishers, the International Committee of Medical Journal Editors (ICMJE) most recent release of “Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals” (International Committee of Medical Journal Editors (ICMJE) 2024) outlines the following requirements related to clinical trial data transparency: First, all clinical trials must be prospectively registered in a publicly accessible registry before enrolment of the first participant. This helps prevent selective reporting and ensures that all trials, regardless of outcome, are documented. Secondly, authors are encouraged to share their data to promote transparency and reproducibility of research findings. This includes making raw data, analysis scripts, and other relevant materials available upon request or through public repositories. Lastly, authors are required to include a data sharing statement in their manuscripts, indicating whether and how they will share data related to the study. This statement helps readers understand the availability of data and facilitates access for further analysis or replication.

Therefore, data-sharing activities in the area of clinical trials, which in the past were considered a burdensome and bureaucratic exercise, diverting scientist from actual research or, even worse, posing a threat to future work due to plagiarism, have become critical and essential components for enabling new research and maximising scientific endeavours.(Cancer Research UK)

However, in the vast majority of cases, clinical trials involve handling participants' personal data throughout the entirety of their execution process, from the recruitment of the first participant to all the way to archiving, and including the sharing of IPD for secondary research purposes. Therefore, it is important to understand how clinical trials participants’ personal data is subject to regulatory frameworks set by national governments.

2.4 Personal Data protection regulation

Personal data is any information that relates to an identified or identifiable individual. This can include a wide range of identifiers, such as name, contact details, identification numbers, online identifiers, and factors specific to the individual's physical, physiological, genetic, mental, economic, cultural, or social identity(Finck and Pallas 2020). Personal data is subject to data protection regulations, which aim to safeguard individuals' privacy and ensure responsible handling and processing of their information (Custers, Sears et al. 2019).

Personal data protection regulation refers to laws and regulations that govern the collection, processing, storage, and sharing of personal data to ensure the privacy and security of individuals' information. These regulations are designed to protect the rights of individuals and establish guidelines for organisations that handle personal data, including businesses, government agencies, CTUs and other entities.

While there isn't a single, global personal data protection guideline akin to the ICH (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use) guidelines for clinical trials, there are several regulations and frameworks that address personal data protection. Examples of notable personal data protection regulations and frameworks include:

- General Data Protection Regulation (GDPR): Enforced by the European Union (EU), GDPR sets out comprehensive requirements for the protection of personal data of individuals within the EU and European Economic Area (EEA). It applies to organisations worldwide that process the personal data of EU/EEA residents.
- UK Data Protection Act 2018 (The National Archives 2018), which incorporates the General Data Protection Regulation (GDPR) into UK law. The Data Protection Act 2018 outlines the rules for processing personal data in the UK

- Personal Information Protection Law (PIPL): Enacted by China, PIPL governs the processing of personal information and imposes requirements on organisations that collect, process, or transfer personal information of individuals in China
- Asia-Pacific Economic Cooperation (APEC) Privacy Framework: APEC developed a privacy framework that provides a set of principles and implementation guidelines for protecting personal information in the Asia-Pacific region.
- ISO/IEC 27701: This is an international standard for privacy information management that provides guidance for organisations on protecting privacy and complying with relevant data protection regulations, including GDPR.

In the United States, data protection regulations are primarily governed by a combination of federal and state laws, as well as industry-specific regulations. Unlike the European Union's comprehensive GDPR, the United States does not have a single overarching data protection law at the federal level. Instead, data protection in the US is regulated by a patchwork of laws that address specific aspects of data privacy and security. For this research, the Health Insurance Portability and Accountability Act (HIPAA) holds special importance as it regulates the use and disclosure of protected health information (PHI) by healthcare providers, health plans, and other covered entities, as well as their business associates.

While these frameworks may not be identical in scope and requirements, they share common principles such as transparency, accountability, and individuals' rights to control their personal data. Organisations operating globally or conducting international clinical trials often need to navigate and comply with multiple regulations to ensure the protection of personal data.

Key principles and requirements commonly found in personal data protection regulations encompass several fundamental aspects. Firstly, there is the principle of lawfulness, fairness, and transparency (Information Commissioner's Office (ICO) 2023), which mandates that personal data must be processed in a lawful, fair, and transparent manner, with individuals duly informed about the collection, usage, and sharing of their data. Additionally, the principle of purpose limitation (Information Commissioner's Office (ICO) 2023) emphasises that personal data should only be collected for specified, explicit, and legitimate purposes, without further processing beyond these objectives. Moreover, data minimisation (Information Commissioner's Office (ICO) 2023) underscores the importance of collecting and processing only the minimum necessary personal data for the intended purpose. Ensuring accuracy (Information Commissioner's Office (ICO) 2023) is also crucial, with personal data required to be accurate and kept up to date, with corrections made promptly when necessary. Furthermore, regulations stipulate storage limitation (Information Commissioner's Office (ICO) 2023), indicating that personal data should only be retained for as long as necessary for the purposes for which it was processed. Security (Information Commissioner's Office (ICO) 2023) is paramount, with requirements for maintaining the integrity and confidentiality of personal data, guarding against unauthorised processing, and preventing loss, destruction, or damage. Lastly, accountability (Information Commissioner's Office (ICO) 2023) is emphasised, with data controllers bearing the responsibility of compliance with data protection regulations and being able to demonstrate adherence to the principles and requirements outlined therein.

In addition to these general principles, personal data protection regulations often include specific provisions related to consent, data subject rights (such as the right to access, rectification, erasure, and portability of personal data), data breaches, cross-border data transfers, and the appointment of data protection officers.

These principles help define the 'Lawful basis' (Information Commissioner's Office (ICO) 2023), which are the legal grounds or justifications that organisations must have to process and share personal data . These include:

- Consent: The individual has given clear consent for their personal data to be processed for a specific purpose.
- Contract: Processing is necessary for the performance of a contract to which the individual is a party, or for taking steps at the request of the individual prior to entering into a contract.
- Legal obligation: Processing is necessary for compliance with a legal obligation to which the data controller is subject.
- Vital interests: Processing is necessary to protect the vital interests of the individual or another natural person.
- Public task: Processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the data controller.
- Legitimate interests: Processing is necessary for the legitimate interests pursued by the data controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject requiring protection of personal data, particularly where the data subject is a child.

A single basis is not inherently better or more important than any other. Organisations must determine and document the most suitable lawful basis for their operation. This determination should be clearly outlined in their privacy policy, where organisations detail how they collect, use, share, and protect personal data. The privacy policy also serves to inform individuals about their rights regarding their personal information.

2.4.1 Enforcement of data protection regulation.

The Information Commissioner's Office (ICO) is the UK's independent body entrusted to uphold information rights in the public interest, promote openness by public bodies and protect data privacy for individuals (Information Commissioner's Office (ICO)). At the supranational level, the European Data Protection Supervisor (EDPS) is responsible, while each EU member state has its own national data protection authority tasked with enforcing pertinent data protection laws and regulations. In the US, the Department of Health & Human Services (HHS) is responsible for data protection and privacy in the healthcare sector.

Compliance with personal data protection regulations is essential for organisations to build trust with their customers, avoid legal liabilities, and mitigate the risks associated with data breaches and privacy violations. Violations of these regulations can result in significant fines, penalties, and reputational damage. Therefore, organisations must implement robust data protection measures and adopt privacy-by-design principles to ensure compliance with applicable regulations.

2.4.2 Privacy protection and CTUs

As handlers of personal data, CTUs are bound by their corresponding personal data protection regulation and must prioritise the protection of their clinical trial participants' privacy. Privacy is a highly complex concept that often encompass various sub-domains such as anonymity (i.e., the ability to hide one's identity), confidentiality (i.e., the ability to share information with a second party without the information being publicly revealed), and solitude (i.e., the right to be left alone) (Figure 2-8).

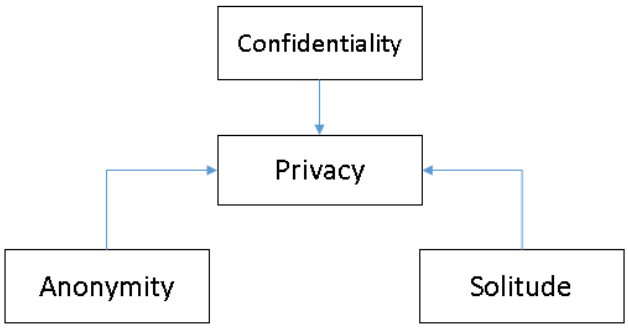


Figure 2-8 Privacy domains

Image source: Original,

Note: Created using the information in Malin et al. 2013 (Malin, El Emam et al. 2013)

Even when a sub-domain is clearly defined, privacy can be very contextual and it is tailored to the circumstances and expectations of individuals. For example, patients' expectations of privacy change when they share information with their doctors versus random people on the street. This expectation could be further modified if the information shared is sensitive or embarrassing. Furthermore, we also need to add a layer that determines what is deemed sensitive from person to person as it can vary(Malin, El Emam et al. 2013). Clinical Trials Units (CTUs) have to deal with all the sub-domains of privacy in their day-to-day operations.

Typically, all data collected in a clinical trial will fall within one of the following four categories in Table 2-2:

id	Type of Variables	Definition
1	Identifier(Health and Services 2003)	Also known as direct identifiers (Tucker, Branson et al. 2016). Data attribute that on its own identifies an individual (e.g., fingerprint) or has been uniquely assigned to an individual. These are well described in the HIPAA "Safe Harbor" release method (see Table 2-4).
2	Indirect identifier(Health and Services 2003)	Also known as quasi-identifiers (El Emam, Jabbouri et al. 2006, Han, Yu et al. 2017). A data attribute that, by itself/on its own, does not identify an individual, but may identify an individual when combined with other information.
3	Special category data (information Commissioner's Office (ICO) 2023)	Also known as Sensitive information(Information Commissioner's Office (ICO) 2012). Data attribute that has a high potential to upset or harm participants wellbeing if the information were to become public.
4	Non-sensitive(Raghunathan 2013)	Data attributes, which are not categorised as direct/indirect identifiers or sensitive information. Such attributes do not need to undergo data protection.

Table 2-2 Types of variables in clinical trials,

Table 2-3 shows examples of the types of variables collected on clinical trials and their classification as per Table 2-2 above

Direct identifier			Indirect identifier			Sensitive information		Non-sensitive Information	
Name	Surname	Mobile number	Gender	Ethnicity	Place of birth	Disease	On Income support	Blood test executed	Height taken
Jane	Smith	05031 407 554	Female	White	London	HIV	No	Yes	Yes
John	Smith	08550 654 757	Male	Black	York	Common cold	Yes	Yes	No
Lucy	Jones	08838 461 454	Female	White	Caracas	Broken leg	Yes	No	Yes
James	Williams	09824 995 422	Male	Black	London	Hepatitis	No	No	Yes

Table 2-3 Examples of variables in clinical trials

Note: This data is fictitious, and it does not represent any living individual.

All data is collected from participants in a confidential manner for clinical trials, after which it is entered in to databases that are often pseudonymised (Ohmann, Banzi et al. 2017) (i.e., each participant is assigned a unique ID number, and if required, it can be

matched with the participant' personal identification details, usually held separately). Subsequently, the clinical trial data, in its corresponding database is processed and/or extracted for its primary use as per the study protocol and analysis plan, and finally, the clinical trial datasets should be suitable prepared for secondary research, including the provision of the necessary documentation for their interpretation and valid use (Modi, Kichenadasse et al. 2023) (Tudur Smith, Hopkins et al. 2015) (Ohmann, Banzi et al. 2017). Data protection regulations apply at every step of this process. For this last step, anonymisation could be a used as a tool to protect participants' privacy and to mitigate some of the obligations imposed by data protection laws (Modi, Kichenadasse et al. 2023) (Tudur Smith, Hopkins et al. 2015) (Ohmann, Banzi et al. 2017).

2.5 Anonymisation

Anonymisation is defined as "the act or process of making anonymous, of hiding or disguising identity" (Wiktionary). However, anonymisation is an evolving concept in relation to data sharing, because it has changed as the technological and regulatory environments progress.

Anonymisation, before the 1990s, used to be about simply removing name and contact details of individuals on the databases. This approach did not make the data anonymous as shown by Sweeney in 1996 (Sweeney 2002), as she successfully re-identified individuals on publicly available and anonymised heath records. This positive re-identification led to the creation and passing of the Health Insurance Portability and Accountability Act (HIPAA) in the US (U.S. Department of Health & Human Services (HHS) 2022), which clearly proposes two methods for data release: "Expert Determination" and the "Safe Harbor" for de-identifying protected health information (PHI). According to HIPAA, de-identified health information is not considered PHI

and is thus not subject to the HIPAA Privacy Rule. The “Safe Harbor” method provides specific criteria for removing identifying information from health data to achieve de-identification and it requires the removal of 18 specified identifiers (or any other unique identifying characteristic) that could potentially link the data to an individual (Table 2-4)

The following are the 18 elements that must be excluded/removed from a dataset:

1. Names;

2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:

    a) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and

    b) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.

3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

4. Telephone numbers;

5. Fax numbers;

6. Electronic mail addresses;

7. Social security numbers;

8. Medical record numbers;

9. Health plan beneficiary numbers;

10. Account numbers;

11. Certificate/license numbers;

12. Vehicle identifiers and serial numbers, including license plate numbers;

13. Device identifiers and serial numbers;

14. Web Universal Resource Locators (URLs);

15. Internet Protocol (IP) address numbers;

16. Biometric identifiers, including finger and voice prints;

17. Full face photographic images and any comparable images; and

18. Any other unique identifying number, characteristic, or code.

When all the above identifiers of the individual or of relatives, employers, or household members of the individual, are removed, the dataset satisfies the de-identification standard as per the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule using the Safe Harbor method (issued by the US Department of Health & Human Services).

Table 2-4 HIPAA privacy rule Safe Harbor de-Identification method

Data Source: The HIPAA Privacy Rule (U.S. Department of Health & Human Services (HHS) 2022)

Note: Some elements from the Safe Harbor De-Identification Method would need to be adapted to fit other regions. For example, in the UK, the ZIP code is not applicable, but a similar measure can be used for the UK postcode. This adaptation ensures that regional specificities in address formatting are taken into account, maintaining the method's effectiveness in protecting individual privacy while still allowing for accurate data anonymisation.

The “Expert Determination” method entails engaging a qualified expert with knowledge and experience to assess whether the risk of re-identification of the data is sufficiently low for the dataset to be considered de-identified and exempt from the HIPAA Privacy Rule. This evaluation considers factors such as the nature of the dataset, the intended use of the data, and the risk mitigation measures in place. However, it's important to note that this assessment can be subjective (National Committee on Vital and Health Statistics 2017).

The certainty and simplicity of “Safe Harbor” makes it very attractive (El Emam 2011) and preferable to the “Expert Determination” method. However, “Safe Harbor” has a number on disadvantages such as the excessive removal of useful information, lack of consideration for longitudinal data, no provisions to deal with “free-text” data, just to cite a few (El Emam 2011).

For the UK and the EU, their corresponding GDPR Recital 26 (PrivazyPlan® 2018) defines anonymous information, as “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”, and anonymised data as “data rendered anonymous in such a way that the data subject is not or no longer identifiable”. While these definitions may appear vague and circular, they emphasise that anonymised data must be stripped of any identifiable information, rendering it impossible to derive insights on a single individual, even by the party that is responsible for the anonymisation. (Table 2-5)

"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."
Table 2-5 UK and EU GDPR recital 26
Data Sources: (PrivazyPlan® 2018) (European Parliament - Council of the European Union 2016) (The National Archives 2016)

In contrast to the HIPAA regulations in the US, the UK and the EU do not have specific mechanisms such as "Safe Harbor" or "Expert Determination" for anonymisation of data. Instead, ICO in the UK and the EDPS (including its associated national enforcing agencies) in the EU have issued general guidance on anonymisation practices (Agencia Española Proteccion Datos and The European Data Protection Supervisor (EDPS) 2021) (Information Commissioner's Office (ICO) 2012) (Information Commissioner's Office (ICO) 2021) (Information Commissioner's Office (ICO) 2022). These guidelines provide recommendations and best practices for organisations to follow when anonymising data to ensure compliance with data protection laws.

When done properly, anonymisation places the processing and storage of personal data outside the scope of the (UK or EU) GDPR or HIPAA by maximising

participants privacy, and in theory, on truly anonymised dataset even data subjects should not be able to identify themselves.

2.5.1 Techniques to achieve Anonymisation

The techniques in Table 2-6 are some general tools recommended (Information Commissioner's Office (ICO) 2012) (Elliot 2016) (Elliot, Mackey et al. 2020) to create anonymised datasets, their main aim is to transform variables to reduce detail, without taking away too much data utility. It is going to be assumed that data has already being cleaned and processed by the time it reaches anonymisation. Cleaning data is not within the scope of anonymisation, its main purviews are privacy protection and maintaining data integrity.

id	Technique Name	Technique Description	Example
1	Perturbation/ Noise addition	Random noise from a known distribution is added to the studied variable. Suitable only for continuous variables and it might reduce the accuracy of the dataset.	The variable date of visit is added a small amount random number of days from a uniform distribution.
2	Swapping	This involves swapping certain data between the records of individuals, making it more difficult to identify individuals by linking together different information relating to them. Also known as shuffling, permutation or randomisation. It can be used if the links between cells in a record are not important). It might reduce the accuracy of the dataset.	The variable height of individuals is scrambled so its values for different records are moved around, so that is no longer connected to other information about that individual.
3	Generalisation	Recoding or banding variables in a deliberate reduction in the precision of data.	Converting a person's exact age into an age range, or a precise location into a less precise location.

id	Technique Name	Technique Description	Example
4	Masking	Consist of changing the characters of a data value, by using a constant symbol (e.g., "*" or "x"). Masking is typically partial, i.e., applied only to some characters in the attribute.	A Post code recorded as AB2 3UZ (street level) will become AB2 XXX (region level)
5	Aggregation	This works like generalisation. Data is aggregated across several records for an attribute. It is only suitable for continuous data. It is also known as micro-aggregation or bucketisation.	The variable age is 30, 35 and 40 for three individuals, this is replaced with 35 (average) for all individuals.
6	Suppression	Omit rows (whole record), attribute (whole variable) or cells (certain combinations of record and variables). Attribute suppression refers to the removal of an entire part of data.	The variable age is 88 years for a single individual, who also when compared with the rest of the group is considerably older. The whole record for the individual can be removed or only the age cell.
7	Pseudonymisation	Attribute values are replaced with pseudonyms on a one-to-one correspondence, this preserves the granularity of the data. It carries similar risks to masking, in that much of the original, unaltered data will be contained in the pseudonymised data, and so data matching techniques might be able to identify individual data subjects. Note: Pseudonymisation should never be considered an effective means of anonymisation, but can be considered a security enhancing measure.	Name and address are replaced with a unique identifier. This is routinely done in clinical trials when entering data onto the databases, (pseudonymisation has the advantage of permitting different records relating to the same individual to be linked without storing direct identifiers in the data, this represents a re-identification risk, but it also can be useful for longitudinal studies).

id	Technique Name	Technique Description	Example
8	Synthetic data(Anbazhagan, Sugumar et al. 2012) (El Emam, Mosquera et al.)	Information artificially manufactured using simulation algorithms to recreate real world data. It can be used for test or validation purposes, and it is an alternative approach to data protection, as it does not pose problems with regards to statistical disclosure control because they do not contain real data but preserve desired statistical properties. (Scottish Longitudinal Study Development & Support Unit 2013)	The Scottish Longitudinal Study Development & Support Unit is currently working on a project that will create synthetic datasets which will resemble the UK Longitudinal Studies to allow researchers to experiment and test ideas before applying to use the real study data, which is under a safe haven model, and also researchers have to be vetted before they are granted access. (Scottish Longitudinal Study Development & Support Unit 2013)
Table 2-6 Anonymisation techniques for individual-level data			
Data Sources ((Sweeney 2002, European Commission 2014),(El Emam and Fineberg 2009, Personal Data Protection Commission Singapore 2018),(Data Protection Commission - Ireland , Information Commissioner's Office (ICO) 2012),(El Emam, Arbuckle et al. 2012)).			

The presented techniques adopt a one-size-fits-all approach to privacy protection. However, it is necessary to further understand their impact on data usability. This will be discussed in section 2.5.4.

2.5.2 Data privacy models for Anonymisation

Several technical data privacy models have been developed by the research community to guarantee and assess data anonymity and to protect datasets from re-identification attacks (Table 2-7).

id	Model name	Model Description	Example
1	k-anonymity(Sweeney 2002)	A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. It prevents record linkage as any given record maps onto at least k other records in the data.	It should be impossible to single out a record from any participant in the dataset because, after anonymisation, all participants belong to a single sub-group. If the smallest sub-group contains 2 participants, then the dataset is considered 2-anonymous.
2	l-diversity(Machanavajjhala, Kifer et al. 2007)	The l-diversity model was designed to handle the drawbacks in the k-anonymity model, which protects the identities to the level of k-individuals but it does not safeguard the individuals corresponding sensitive values. This especially happens when there is homogeneity of sensitive values within a k-group. Therefore, the concept of intra-group diversity of sensitive values is promoted within the anonymisation process.	In a 2-anonymous dataset, if the smallest sub-group (consisting of 2 participants who are both male and aged between 40 and 45 years old) includes members who use recreational drugs, and we also know that John, who is 43 years old, participated in the study, we might conclude that John uses recreational drugs. To protect John's privacy, we need to ensure diversity in the stigmatising attribute within the sub-group.

id	Model name	Model Description	Example
3	t-closeness(Li, Li et al. 2007)	t-closeness of a sub-group is attained when the sensitive attribute distance in this sub-group is not greater than the threshold, t with the attribute distance in the whole table. The table is believed to have t-closeness if all sub groups have t-closeness. It is an improvement to l diversity because it tackles skewness.	In a 2-anonymous 2-diverse dataset, the smallest sub-group consists of 2 participants who are both male and aged between 40 and 45 years old, with all members having HIV. If we know that John, who is 43 years old, is in that study, we might conclude that John has HIV. Additionally, we observe that the incidence of HIV is skewed, and the distribution of the stigmatising parameters in the smallest sub-group should match the distribution in the overall dataset or deviate no more than t.
4	Differential privacy (Dwork 2011)	It is a sophisticated technique that guarantees that even if someone has complete information about all but one person in a dataset, they still cannot deduce the information about the final person. This is achieved by adding random noise to the aggregated data. It makes it possible for tech companies to collect and share aggregate information about user habits, while maintaining the privacy of individual users. It requires lots of resources and it is not applicable to small datasets.	In a web survey, patients are asked if they have HIV. If John clicks "yes," a differential privacy algorithm will flip a coin. If the coin lands heads, his "yes" response will go unaltered into the database. If the coin lands tails, the algorithm will flip a second coin: if it lands heads, "yes" will be sent to the database; if it lands tails, "no" will be sent instead. Under this scenario, we cannot trust individual answers, but because we know the amount of noise introduced by the algorithm, we can accurately calculate the total number of HIV patients.

Table 2-7 Data privacy models for anonymisation

Note: k-anonymity, l-diversity and t-closeness are explicitly mention in the "Opinion 05/2014 on Anonymisation Techniques" by the European Commission (European Commission 2014)

K-anonymity is the most popular, but these methods still have intrinsic weaknesses and datasets could still be re-identified, so the quest for novel methods of

anonymisation continues. This area is evolving at an immense speed, as there is more computational power at our disposal.

2.5.3 Re-identification risks

Even after a dataset is declared anonymised, there remains a risk of re-identification, wherein specific individuals could potentially be linked back to the data (Narayanan and Felten 2014). Re-identification risk arises when external information, such as publicly available data or data from other sources, is combined with the anonymised dataset, allowing individuals to be identified (Sweeney 2013). Advanced data analysis techniques, machine learning algorithms, or even simple cross-referencing with publicly available information can sometimes enable re-identification.

Re-identification risk poses a significant concern in data privacy and protection. Therefore, anonymised datasets should aim to carry only a minimal and acceptable risk of re-identification (El Emam, Rodgers et al. 2015). The acceptable level of risk is a function of the information inside the dataset and the wider context of its release (El Emam, Rodgers et al. 2015). For example, patients enrolled in a trial investigating the effects of a medicinal product on the common cold virus might not be concerned about others discovering their participation. However, if the investigated disease is the human immunodeficiency virus (HIV), the patients would (rightly so) have higher expectations of privacy.

If the re-identification risk increases or if participants can be re-identified, then that could bring the dataset back into the realm of data protection laws. Furthermore, if a re-identification actually occurs, it will automatically subject the dataset to the jurisdiction of applicable data protection laws and entail legal repercussions to the organisation responsible (defined as data controllers) for the anonymised dataset. Consequently, mitigating re-identification risk often involves implementing strong

anonymisation techniques, as well as adhering to data protection regulations and standards to safeguard individuals' privacy. Anonymised data release thus becomes a critical exercise in risk management, aiming to strike a balance between the greater good of data sharing versus individuals' information rights. El Emam et al. (El Emam, Rodgers et al. 2015) provides a very concise graphical explanation in Figure 2-9.

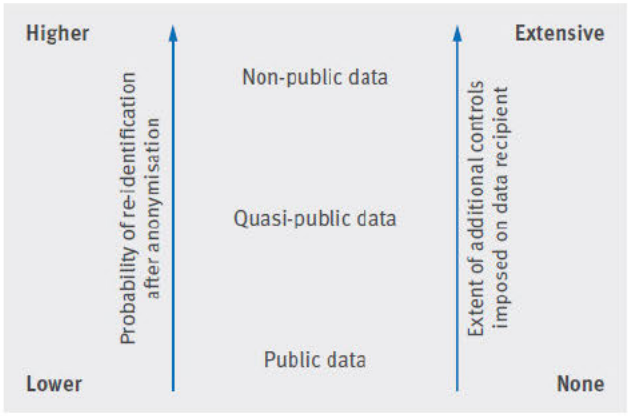


Figure 2-9 Risk of re-identification vs data access controls
Image source: El Emam et al. 2015 (El Emam, Rodgers et al. 2015)

Unequivocally, there will always be circumstances in which any dataset should be shared, and a manner in which it can be shared (Gøtzsche 2011) (Longo and Drazen 2016). If individuals cannot be re-identified then the controls around data release can be minimal or non-existent. However, if the data is hard to robustly anonymise, its access should be controlled (e.g., with a data use agreement, Figure 2-10).

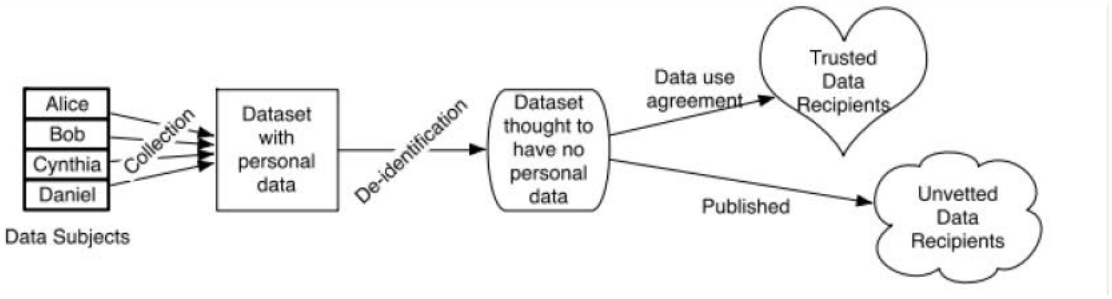


Figure 2-10 Data collection, de-identification and use

Image source: U.S. Department of Commerce, National Institute of Standards and Technology Internal Report 8053 (Garfinkel 2015)

Intuitively, we can conclude that to be compliant with the recommendations by data protection regulators and minimise risk of re-identification, we should provide the highest level of protection to participants, by generating data that is anonymised to the maximum. The problem in this argument is that we would be drastically reducing data utility.

2.5.4 Anonymised data utility

Any step towards anonymisation, regardless of the techniques and/or model used, reduces the original information in the dataset by some extent. It is very hard to completely anonymise data while still leaving it in an analysable and usable form (Tudur Smith, Hopkins et al. 2015). There is an inevitable trade-off between data utility and anonymity, which is shown in Figure 2-11.

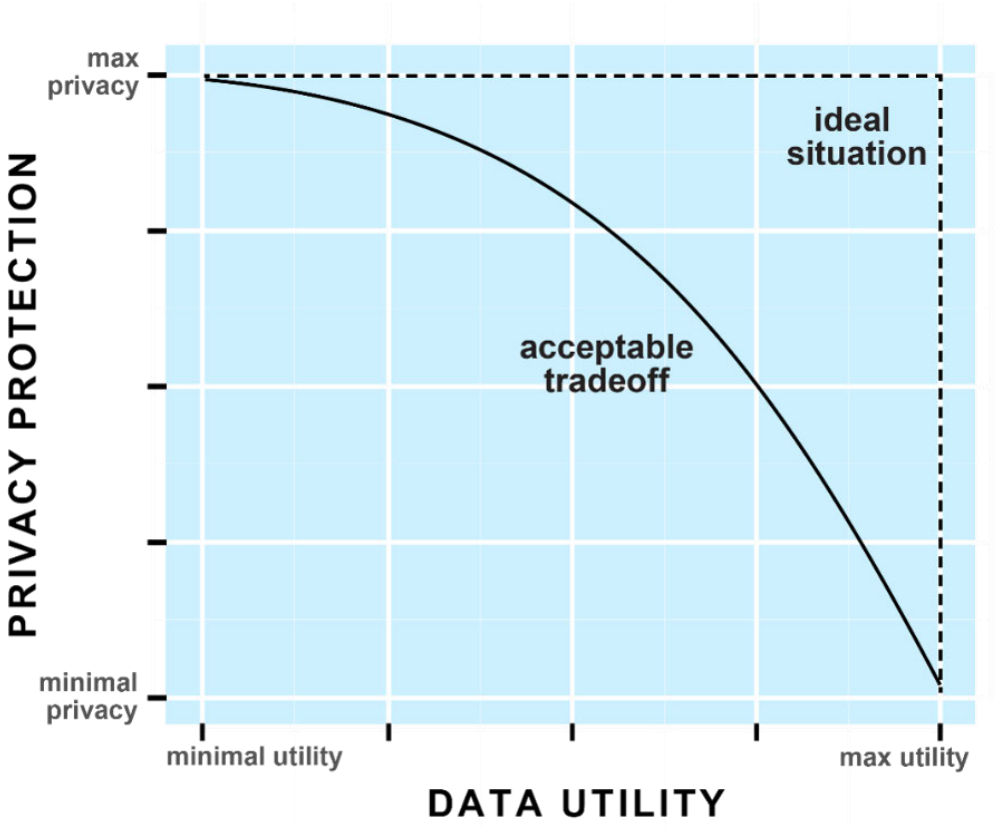


Figure 2-11 The trade-off between perfect data and perfect privacy

Image source: Replica Analytics Ltd, author: Professor Khaled El Emam
https://www.dataguidance.com/opinion/international-what-synthetic-data-and-how-it

Permission given by the author to reproduce this content

Generally, as the amount of anonymisation increases, the utility (e.g., clarity and/or precision) of the dataset is reduced (El Emam et al. 2013)(El Emam and Arbuckle 2013). To strike the perfect balance between data protection and data utility, organisations need to decide in advance the degree of the trade-off between acceptable (or expected) utility and the effort to reduce the risk of re-identification in their datasets. It is necessary to define clearly the objective of the anonymisation, because it should be tailored specifically to the purpose at hand.

2.5.5 Anonymisation and IPDs of clinical trials for sharing

The anonymisation approaches outlined in this chapter are versatile and applicable to any kind of personal dataset. Consequently, they could be employed to facilitate the sharing of individual participant data from clinical trials. By using these approaches, researchers and organisations (including CTUs), could share data more freely without compromising the confidentiality of participants, thus safeguarding their privacy. However, the specific requirements for a particular situation may be complex. The interplay between levels of anonymisation, data utility, and legal requirements, combined with the resources available, may prevent sharing at an ideal level. The approaches presented in this chapter promote transparency, facilitate collaboration, and support the advancement of medical research, while adhering to ethical standards and data protection regulations. However, understanding how these approaches are tailored to clinical trials is essential, and this is further explored in the next chapter.

2.6 Conclusion

Executing clinical trials demands a significant amount of effort, as they require meticulous planning, coordination, and resources. Clinical trial units (CTUs) play a crucial role in this process, assembling all the necessary components to conduct these trials (see Figure 2-12), ultimately producing documentation, results, and data. The Individual Participant Data (IPD) obtained at the end of the study, which is very expensive to acquired, holds immense value beyond the initial research question, serving various purposes that are increasingly recognised through requests for clinical trial data transparency.

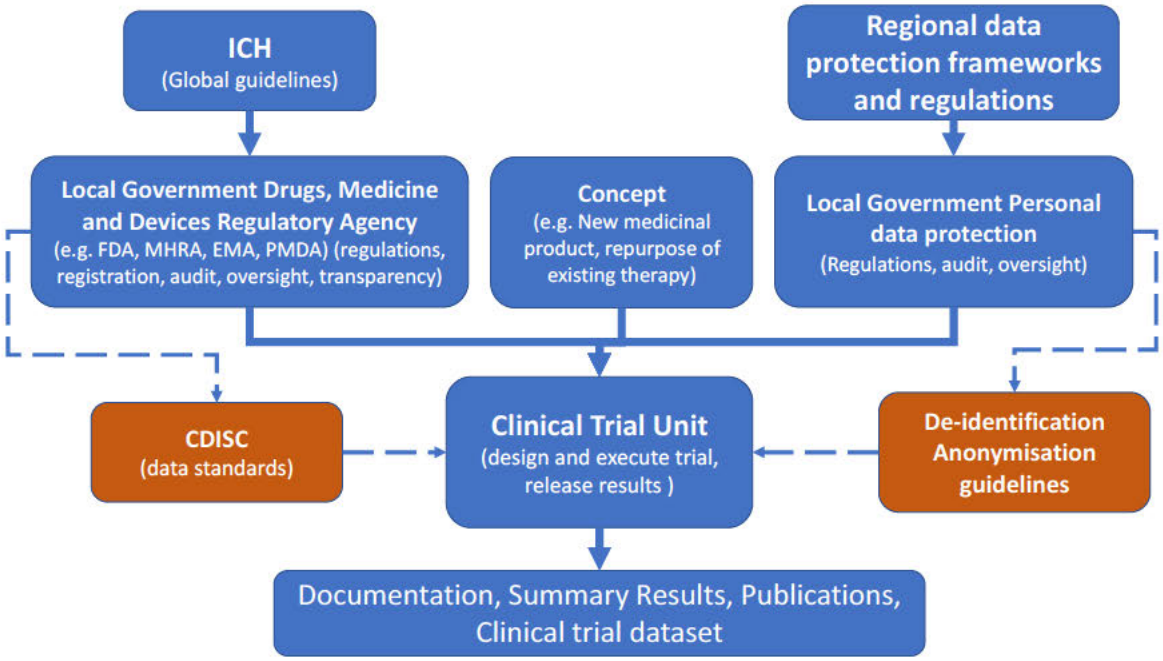


Figure 2-12 Clinical trials output generated by Clinical Trial Units

Image source: Original

Note: Dashed lines represent optional relationships, while solid lines represent mandatory ones. Created with information from this Chapter.

However, IPD is protected by data protection regulation such as GDPR (respectively in the UK and EU) and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) in the US. These regulations are designed to safeguard personal data, including IPD, by restricting its sharing and processing.

For example, under GDPR, personal data cannot be shared freely unless specific conditions are met, such as obtaining explicit consent from individuals or ensuring that the data is anonymised to remove any identifying information.

To this end, data protection regulators have provided valuable guidance in anonymisation. However, this guidance often follows a one size-fits-all approach, applying uniformly across all kinds of organisations (Information Commissioner's Office (ICO) 2012). This means that routinely collected data, marketing data, genomic data and clinical trials datasets are all treated under the same umbrella, despite each dataset having its own particular challenges.

It is important to note that privacy protection addressing the needs on routinely collected data and big datasets seems to be well-researched, generating numerous useful anonymisation techniques and models. These techniques could potentially be applicable to small datasets, such as those derived from clinical trials, as they provide an additional layer of protection. Moreover, they could serve as valuable tools for estimating the risk of re-identification.

With that said, anonymised clinical trials datasets have their own hurdles to clear before they can be shared, as acknowledged by others (Hughes, Wells et al. 2014) (Narayanan and Felten 2014) (El Emam, Rodgers et al. 2015). These challenges include for example, the presence of many indirect identifiers needed for data analysis or dealing with "free text verbatim terms". Therefore, in the next chapter, we further delve into the current recommendations and practices for anonymising IPD from clinical trials to facilitate sharing. This marks the first piece of original research for this project.

Chapter 3 Scoping review

3.1. Introduction

This chapter relates to the first objective of the PhD: to describe the available anonymisation methods and techniques for clinical trials datasets.

As discussed in the previous chapter, various techniques and data privacy models exist for achieving anonymisation, but many are general and applicable to any kind of dataset. Therefore, we aimed to understand the specific recommendation and methods available for anonymising clinical trial datasets.

To this end, I executed a scoping review (Peterson, Pearce et al. 2017) (Peters, Godfrey et al. 2015) because it provides a broad overview of a research area, clarify concepts, identify gaps and synthesises diverse evidence. I used the Joanna Briggs Institute (JBI) methodology(The Joanna Briggs Institute 2015) which outlines six steps as follows: 1) identifying the research question; (2) identifying relevant studies; (3) study selection; (4) charting the data, (5) collating, summarising, and reporting the results; and (6) (optional) consultations.

The objective of this scoping review was to describe and compile all available ideas and recommendations on anonymisation of clinical trial datasets in one comprehensive place.

Steff Lewis, Christopher Weir and Aryelly Rodriguez identified four indicator papers (Hrynaszkiewicz, Norton et al. 2010, Tudur Smith, Hopkins et al. 2015, Ohmann, Banzi et al. 2017, Keerie, Tuck et al. 2018) that provided a robust foundation of the concepts of anonymisation in clinical trials datasets. These papers were used to design the search strategy with the UoE librarian, Ms Marshall Dozier, whose input was so critical that she qualified to be a co-investigator on this part of the research.

With the objective and search strategy defined, I initiated the drafting of a systematic scoping review protocol, which underwent peer review by my co-authors. The protocol was finalised on the 11th Jan 2019. Subsequently, I spearheaded the execution of the protocol, with contributions from my co-authors as follows: Chris Tuck and Alastair Murray served as second reviewers for screening titles/abstracts and full documents, the remaining co-authors arbitrated any disagreements.

Since this part of the research did not use any participant data, a full ethics application was not required. I only had to complete a "Self-Audit Checklist for Level 1 Ethical Review" (Appendix 1) to verify that there were no foreseeable ethical risks.

To the best of our knowledge, this was the first scoping review on anonymisation methods/techniques for clinical trials datasets.

After completing all pre-defined protocol activities in May 2021, I compiled a manuscript for publication. I carefully addressed all comments from my co-authors before submitting it for publication on the 9th Jun 2021. Additionally, I managed and addressed the journal peer reviewers' and editorial comments until the manuscript was accepted for publication on the 11th Feb 2022.

The scoping review was published as a peer-reviewed paper in the Clinical Trials SAGE Journal on the 22nd June 2022 and is openly licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Our article, titled "Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review" (Rodriguez, Tuck et al. 2022), is available at https://doi.org/10.1177/17407745221087469 and is presented in the next item of this chapter. All additional published material related to this article is included in Appendix 2 (including the protocol).

3.2. Published article

Review

CLINICAL TRIALS

Clinical Trials
2022, Vol. 19(4) 452–463


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17407745221087469
journals.sagepub.com/home/ctj
SAGE

Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Aryelly Rodriguez[1], Christopher Tuck[2], Marshall F Dozier[3], Stephanie C Lewis[1], Sandra Eldridge[4], Tracy Jackson[5], Alastair Murray[6] and Christopher J Weir[1]

Abstract

Background/Aims: There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community, and differing recommendations exist on how to perform anonymisation prior to sharing. We aimed to systematically identify, describe and synthesise existing recommendations for anonymising clinical trial datasets to prepare for data sharing.

Methods: We systematically searched MEDLINE®, EMBASE and Web of Science from inception to 8 February 2021. We also searched other resources to ensure the comprehensiveness of our search. Any publication reporting recommendations on anonymisation to enable data sharing from clinical trials was included. Two reviewers independently screened titles, abstracts and full text for eligibility. One reviewer extracted data from included papers using thematic synthesis, which then was sense-checked by a second reviewer. Results were summarised by narrative analysis.

Results: Fifty-nine articles (from 43 studies) were eligible for inclusion. Three distinct themes are emerging: anonymisation, de-identification and pseudonymisation. The most commonly used anonymisation techniques are: removal of direct patient identifiers; and careful evaluation and modification of indirect identifiers to minimise the risk of identification. Anonymised datasets joined with controlled access was the preferred method for data sharing.

Conclusions: There is no single standardised set of recommendations on how to anonymise clinical trial datasets for sharing. However, this systematic review shows a developing consensus on techniques used to achieve anonymisation. Researchers in clinical trials still consider that anonymisation techniques by themselves are insufficient to protect patient privacy, and they need to be paired with controlled access.

Keywords

Clinical trials, systematic review, data anonymisation, patient identification systems, personally identifiable information, datasets, data curation, guidelines

[1]Edinburgh Clinical Trials Unit, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
[2]Centre for Cardiovascular Science, The University of Edinburgh, Edinburgh, UK
[3]Library & University Collections, Information Services, The University of Edinburgh, Edinburgh, UK
[4]Pragmatic Clinical Trials Unit, Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK
[5]Asthma UK Centre for Applied Research, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK
[6]Independent Researcher, Edinburgh, UK

Corresponding author:
Aryelly Rodriguez, Edinburgh Clinical Trials Unit, Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Level 2, Nine Edinburgh BioQuarter, 9 Little France Road, Edinburgh EH16 4UX, UK.
Email: aryelly.rodriguez@ed.ac.uk

Introduction

Clinical trials are complex, time-consuming and costly, and it is wasteful not to use data fully.[1] Therefore, when academic-led clinical trials are completed, their results are usually released to the public and wider scientific community in scientific journals or clinical trials registries. Existing clinical trials' data can be used to answer novel clinical questions, to reproduce and check analysis, to understand basic science, to investigate new methodologies and for teaching.[2] Also, there are sometimes considerable amounts of data that are not analysed as part of the published results.[3] In addition, trial data are often useful after the end of a trial to perform meta-analyses across several trials and using the individual patient data from each trial adds to the quality of such analyses,[4] for instance, by allowing full investigation of subgroup effects. There is now a drive, particularly from publishers and funders, to encourage the general release of relevant anonymised trial datasets[5] among interested parties.

Clinical trial datasets contain personal health information of the trial participants. It is imperative that data sharing does not disclose personal data to anyone who falls outside the original group to whom the trial participants have provided consent to access their data. Anonymising the trial dataset fulfils this requirement. However, the anonymisation process removes information from the data, and if not done carefully, the original trial analyses could not be reproduced, which in turn will limit the data's usability for further research.[6]

The drive to share data more widely has generated various sets of recommendations to enable sharing.[5,7–10] Embedded within these, there is a variety of recommendations on how to anonymise a dataset.

Why it is important to do this review

To our knowledge, there are no reviews of the methods and/or recommendations for the process of generating anonymised clinical trial datasets (a search was executed on the 15 February 2021 on Google Scholar[11] with 'literature' 'review' 'anonymization' 'methods' 'clinical trials' and also 'literature' 'review' 'anonymisation' 'methods' 'clinical trials', the first 100 results were screened for each search and relevant results were not found).

To understand and collate the techniques used or recommended for data anonymisation in clinical trials, a systematic scoping review is required.

Objective

To identify, describe and synthesise the existing methods/recommendations to anonymise datasets from clinical trials.

Methods

The *Joanna Briggs Institute Reviewers' Manual: 2015 Methodology for JBI Scoping Reviews*[12,13] and the PRISMA Extension for Scoping Reviews (PRISMA-ScR)[14] were followed for the execution of this scoping review.

Types of publications

We included any publications or documentation giving recommendations on anonymising datasets from clinical trials in any therapeutic area. Non-empirical publications, such as editorials, expert views or practice guidelines were also included in this review

Type of outcomes

The primary outcome is the reported methods and/or recommendations for anonymisation of clinical trials datasets.

Search methods for identification of publications

We performed a comprehensive systematic search to identify publications reporting methods or recommendations for anonymising clinical trials datasets. No language restrictions were imposed to attempt worldwide coverage. We did not identify any non-English publications.

Electronic searches. Web of Science (WoS), MEDLINE® (including non-indexed and in-process records) and EMBASE databases were searched from inception to 11 February 2019. The searches were rerun from 1 January 2019 to 8 February 2021 for MEDLINE® and EMBASE. A discrepancy was identified by M.F.D. in the original WoS strategy, so that, we reran the complete search from inception to 8 February 2021.

The search strategy used the following key concept areas, adopting subject headings and keywords as relevant for each database:

(Clinical) and
(trial* or randomi* or research* or control*) and
(principle* or guid* or recomm*) and
(shar* or reus* or re-us* or access* or open) and
(de-identi* or deidenti* or anonym* or privacy or confidential*)

The search was piloted with four indicator papers (Ohmann,[5] Keerie,[9] Tudor-Smith[15] and Hrynaszkiewicz[16]) that the searches needed to retrieved to ensure their effectiveness. The resulting detailed electronic search strategies are presented in Appendix 2 in the supplemental materials.

Searching other resources. To ensure the comprehensiveness of our search, we searched the websites of major research governance organisations and public research funding bodies as recommended by the Health Research Authority[17] and the Wellcome Trust,[18] the top 10 wealthiest charities,[19] the top 10 UK charities by brand value[20] and all registered UK academic clinical trials units,[21] to find guidelines published as grey literature from February 2019 until March 2020, so as not to omit documents not published as journal articles and not indexed in the bibliographic databases.

To further supplement our search field, we used citation and reference tracking (backwards and forward citation searching) on the selected articles from the electronic searches in order to identify additional sources. Preliminary results of this project were presented at the Fifth International Clinical Trials Methodology Conference 2019[22] where we requested to be contacted by any author or expert who could assist with the project but we did not receive any replies. During this event, several colleagues suggested publications to include in our grey literature.[23,24] Shortly after, the COVID-19 pandemic started and we decided not to burden authors/experts with our requests and to concentrate on getting this project executed with the evidence that we had already collected. All the items included in this review obtained via the search of other resources were re-checked on May 2021 to locate updated versions since the original search.

Data collection and analysis

Records were retrieved and transferred into the reference manager EndNote,[25] which was used for de-duplication and to maintain a master library of the records throughout the review process. Covidence software[26] was used for further de-duplication, screening and full-text review. Two reviewers (A.R. and either C.T. or A.M.) independently screened titles and abstracts for eligibility. Full-text copies of all potentially relevant records were obtained using the reference manager.

Records identified from citation and reference tracking, and major research governance organisations, public research funding bodies and charity websites were collated in MS Excel,[27] for manual de-duplication and title screening. Records selected for full-text review were manually retrieved. Two teams (A.R. and either C.T. or A.M.) independently assessed whether each full-text record met the inclusion criteria. Chosen full-text records were added to the master library in EndNote.[25]

Any discrepancies were discussed between the reviewers and if agreement could not be reached then it was arbitrated by a third reviewer (S.C.L., C.J.W. or S.E.).

Publications were excluded if they did not have concrete recommendations/methods of anonymisation, or they were not from a clinical trial framework, or they were focused on omics data or big data.

Data extraction/management and synthesis. A data extraction form to collect relevant data items from eligible sources was developed and piloted in line with Cochrane guidance,[28] this included: publication details (Authors names, Journal, year), country and classification (from electronic search or from other sources).

Data extraction and analysis was undertaken by one reviewer (A.R.) in NVivo®[29] using thematic synthesis.[30,31] Therefore, the included records were read 'line-by-line', and when recommendations/methods on anonymisation were found, they were coded to a theme. At this stage, we allowed themes to be free and data-driven (i.e. to emerge from the data), rather than rigidly defining them a priori. It was possible to assign several themes to the same sentence. An independent sense-check was conducted by a second reviewer (A.M.) of the free themes. Any discrepancies were discussed between the reviewers and if an agreement could not be reached then it was resolved by a third reviewer (S.C.L., C.J.W. or S.E.).

The free themes were grouped into broader themes by the study team, this was repeated until we reached a final theme structure. We did not attempt to generate analytical themes[30] as our goal was to only identify the existing recommendations/methods on anonymisation.

Finally, the data from the included publications were summarised in descriptive tables. Themes were summarised by narrative analysis[32] and if applicable descriptive statistics.

Results

We identified 1059 potentially eligible records (Figure 1 in the online supplemental materials). Six hundred thirty-seven records were excluded after title and abstract screening. Three hundred sixty-three records were excluded after full-text review. Fifty-nine records[5,9,15,16,23,24,33–86] (representing 43 studies) met the inclusion criteria and were included in the final qualitative synthesis (Appendix 3 has the full list and characteristics of the included records).

Included studies' characteristics

Table 1 summarises the observed characteristics of the included studies and their associated records, it also shows the included studies by source and country/region and year of publication. Figure 2 in the online supplemental materials shows the included studies over time.

Table 1. Studies/record characteristics.[a]

Parameter	Category	Studies *N* = 43, *n* (%)	Records *N* = 59 *n*(%)
Source[b]	Electronic search	19 (44)	21 (36)[c]
	Other sources	24 (56)	38 (64)[d]
Country/region	EU	12 (28)	24 (39)
	UK	11 (26)	14 (23)
	US	10 (23)	12 (20)
	Canada	5 (12)	5 (8)
	Australia	2 (5)	2 (3)
	US–EU–UK	2 (5)	3 (5)
	South Korea	1 (2)	1 (2)
Year of publication	2003–2008[b]	5 (12)	5 (8)
	2009–2014[b]	15 (35)	17 (29)
	2015–2020[b]	23 (53)	37 (63)
Studies split by source			
Parameter	Category	Electronic search *N*=19, *n* (%)	Other sources *N*=24 *n*(%)
Studies split by country/region	Canada/US	6 (32)	9 (37)
	EU/UK	12 (63)	11 (46)
	Other regions[e]	1 (5)	4 (17)
Studies split by year of publication	2003–2008[f]	3 (16)	2 (8)
	2009–2014[f]	7 (37)	8 (33)
	2015–2020[f]	9 (47)	14 (58)

[a]Therapeutic field was not applicable and it was not recorded.
[b]Where applicable, the oldest record in the included study determined the overall study date.
[c]Corresponding references[5,9,15,16,33–46,48,49,87]
[d]Corresponding references[23,24,50–84,86]
[e]Consisting of Australia, the United States–EU–the United Kingdom and South Korea.
[f]Where applicable, the oldest record in the included study determined the overall study date.

Deriving the coding themes

A NVivo® exploratory word cloud was generated, it displayed the frequency in which significant words appeared in the included studies from the electronic searches (Figure 3 in the online supplemental materials), and it provided an initial idea of the themes present in the available data.

A.R. started the coding into free themes. As the actual coding progressed, the themes were reviewed and grouped by the study team until its structure was locked on 5 September 2019 by A.R., S.C.L. and C.J.W. The subsequent coding of the studies from other sources did not add any new themes. Eleven themes were identified (see Table 1 in the online supplemental materials).

The body of knowledge after coding themes

The 11 themes were applied to all 43 included studies (see Table 2). The most common theme among the selected studies were the definitions of de-identification (34 studies (79%)), anonymisation (28 studies (65%)), techniques for the manipulations of data (34 studies (79%)) and the implementation of controlled access for data release (38 studies (88%)).

In general terms, when study authors described anonymisation, de-identification and pseudonymisation, their explanations gravitated around the definitions presented in Table 3.

The described aim of data manipulation is to transform variables to reduce detail, without taking away too much data utility. The most common data manipulation methods' definitions are given in Table 3.

Twelve studies (28%) recommended the use of privacy models (such as k-anonymity,[88] l-diversity[89] and differential privacy[90]) to further guarantee and assess data anonymity to protect datasets from re-identification attacks.

The theme of controlled access mostly referred to the implementation of data-sharing agreements, the location of data behind a secure access barrier (either physical, virtual or both), the identification and vetoing of secondary research (e.g. checking requesters are bona fide researchers with a valid research question). In contrast, the theme of open access referred to minimal (or non-existent) requirements for allowing access to the data set to secondary researchers.

Central repositories (mentioned by five studies (12%)) were described as destinations where institutions

Table 2. Themes by studies.

Id/theme	Studies (*N* = 43)		Associated records
	n	%	
1. Anonymisation	28	65	5, 15, 16, 33, 34, 36, 37, 42, 44–48, 50–54, 56, 57, 63, 65, 66, 68–71, 77–80, 82, 83, 85–87
2. De-identification	34	79	5, 15, 16, 33, 34, 36, 37, 42, 44–48, 50–54, 56, 57, 63, 65, 66, 68–71, 77–80, 82, 83, 85–87
2.1. HIPAA identifiers[a]	23	53	5, 16, 24, 33, 34, 36–39, 43, 46, 47, 52–54, 56, 57, 59, 60, 63, 65, 66, 76, 77, 80, 82–87
2.2. Hrynaszkiewicz identifiers[b]	12	28	9, 16, 33, 45–47, 59, 60, 63, 66, 74, 78, 85–87
3. Pseudonymisation	23	53	5, 9, 34, 36, 37, 40–44, 46, 49–51, 55–57, 66, 68, 69, 71, 77, 82
4. Manipulation of data	34	79	9, 15, 16, 24, 33, 36–38, 42–48, 50, 53, 54, 56, 57, 59, 61, 62, 64–66, 69, 71, 75, 77, 78, 80–84, 87
4.1. Perturbation[c]	7	16	9, 36, 55, 66, 67, 77, 84
4.2. Recalculation[c]	12	28	9, 16, 23, 43, 45, 52, 54, 56, 59, 63, 64, 67, 70, 73, 78, 80, 82, 83
4.3. Recoding[c]	16	37	9, 33, 35, 43, 51–55, 59, 60, 63, 64, 66, 67, 69, 70, 72, 77–79, 82–84
4.4. Suppression[c]	17	39	9, 35, 45, 51–54, 56, 57, 59, 60, 62, 63, 65–67, 69, 70, 72, 73, 77, 78, 80, 82–84
4.5. Remove superfluous data[c]	2	5	45, 48
5. Privacy model	12	28	35, 38, 40–42, 46, 55, 57, 66, 69, 84–86
5.1. K-anonymity[c]	7	16	35, 38, 40, 55, 57, 69, 84
6. Controlled access	38	88	5, 9, 15, 16, 24, 33, 34, 36–39, 44–48, 50, 54, 57, 59, 60, 62–66, 71, 72, 75, 77–87
6.1. Black box[c]	3	7	41, 43, 46
6.2. Encryption[c]	8	19	36, 39–42, 57, 66, 77
6.3. Safe haven[c]	8	19	33, 36, 43, 46, 47, 55, 66, 83, 87
6.4. Split location[c]	5	12	34, 41, 43, 66, 81
7. Open access	7	16	9, 15, 36, 50, 56, 66, 85, 86
8. Central repositories	5	12	16, 33, 46, 62, 66
9. Expert determination	12	28	16, 24, 38, 46, 51, 66, 65, 74, 76, 77, 80, 82–84
10. Provision of context documents	12	28	5, 9, 15, 44, 46–48, 62–64, 66, 75, 79, 82, 83, 87
11. Risk calculation	15	35	16, 33, 35–37, 47–49, 57–59, 66, 69, 72, 78, 81, 82, 84–87

HIPAA: Health Insurance Portability and Accountability Act.
[a]HIPAA identifiers refers to the HIPAA Safe Harbor method that requires the removal of 18 items of protected health information.[76]
[b]Hrynaszkiewicz identifiers refers to the removal of direct identifiers (information sources such as name and/or address, which on their own can re-identify participants) and the consideration/removal of indirect identifiers (variables that on their own might not represent a risk of re-identification for participants but in combination with of other indirect identifiers might increase the risk of re-identification, e.g. sex combined with age).[16]
[c]These are child codes that are included in their parent code.

could deposit their datasets to be managed by a third party and accessed by secondary researchers.[92–95]

The expert determination method for dataset release (12 studies (28%)) was generally described as when an expert (chosen for their knowledge/qualification) could assess the risk of re-identification of clinical trial datasets using 'generally accepted statistical and scientific principles',[66] if the risk is low, the data are certified and granted release to a secondary researcher.

Twelve studies (28%) recommended the provision of documental context to avoid erroneous interpretation and use of the anonymised datasets. Suggested documents to be provided included: original study protocol (and applicable amendments), statistical analysis plan, annotated case report forms and a data dictionary.

Finally, 15 studies (35%) highlighted the importance of assessing the risk of the anonymised dataset before making a decision on release, however, only four records[35,59,66,69] (three studies) described how the risk could be calculated.

Most suggested processes for sharing anonymised datasets

Thirty-five studies (81%) described that at the end of a clinical trial, data should be de-identified (key items stripped from the dataset). Following this, data manipulation techniques should be used to further anonymise the datasets. Finally, the datasets should be made available under a controlled access approach.

Thirteen of those 35 studies also mentioned a step before release under controlled access in which the risk of re-identification should be assessed. This would start an iterative process, and once the risk is deemed acceptable, the anonymised data set should be made available under controlled access (Figure 1).

Discussion

The EU/UK region provided 53% of the included studies, followed by the US/Canada region with 35%, while

Table 3. Most common definitions for anonymisation, de-identification and pseudonymisation.

Pseudonymisation	De-identification	Anonymisation
• Attributes are replaced with pseudonyms on a one-to-one correspondence • It is never an effective means of anonymisation • A security enhancing measure • Pseudonyms bear no relation to the patient details • Preferably reversible	• Stripping datasets of patients identifying variables as per either: ○ HIPAA 18 items 'Safe Harbor' method (US) ○ Hrynaszkiewicz et al. 28 items of personal and clinical information (Europe)	• Any given record lacks any individuality, distinction or recognisability • Can potentially distort data • The link with the original dataset should be destroyed • Set at a level to reach acceptable risk, but binary in law

Most common definitions for data manipulation techniques[a]

Suppression (removal, elimination)	Recoding (grouping, masking, replacement, generalisation, blurring, aggregation)	Recalculation	Perturbation
• Delete outliers • Delete free-text • Delete high-risk variables • Delete high-risk records	• Keep first three digits of postcode • Categorise age (18–40) and >= 40	• Show age instead of DOB • Show study day relative to randomisation day, instead of date (e.g. day 7) • When dates are important they are presented offset	• Add random noise to variables • Replace data with simulated random values • Data shuffling • Rounding of variables

HIPAA: Health Insurance Portability and Accountability Act.
[a]Tuck et al.[45] and Tudur-Smith et al.[48] mentioned the removal of superfluous data (e.g. deletion of data, such as audit trails) to supplement data manipulation techniques.

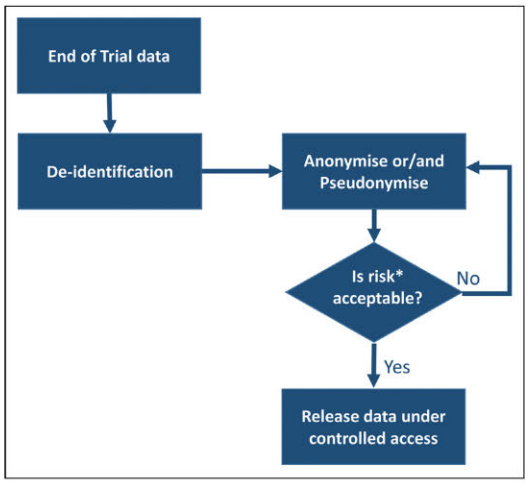


Figure 1. Most suggested method to release anonymised datasets from clinical trials.
Risk of re-identification is a complex variable, which is minimised using controlled access. The description of risk is out of scope for this review. Other processes: five studies described usage of open access instead, one study mentioned both controlled and open access for data release and the remaining two studies did not discussed data release.

the rest (12%) originated from other regions. This result was very similar when studies were split by source. Similarly, 53% of the included studies were published after 2015, 35% of the studies were published from 2009 to 2014 and the rest (12%) of the studies were published from 2003 to 2008. This profile was also observed when the studies were split by source, this shows the greater interest in this topic as time progresses. Overall, the EU/UK region from 2015 to 2020 was the most prolific with 16 studies out of 43 (37%). Where the content in the included studies was congruent regarding the source of the studies, this was noted, while the studies from other sources were coded because there was no need to update the coding themes generated with the studies from the electronic searches. However, a small but crucial difference is that studies from other sources have more detail and examples regarding data manipulations; this is most probably due to the lack of restriction on publication size for this type of source.

Topic 1: The relationship among the themes, pseudonymisation, de-identification and anonymisation, in the context of clinical trials. Anonymisation versus de-identification: they are both described as tools to facilitate data sharing. They rarely appear in isolation in any of the included studies, because they are part of the wider theme of data transparency and patient privacy. In this review, seven records coded to anonymisation, 14 records to de-identification and 28 records coded to both themes.

Anonymisation is presented as an abstract theme with lots of interpretation, mostly shaped by the regional laws where the publications originated (i.e. each researcher would have a theme that they favour which is shaped by their legal framework). These laws could be vague with their definitions and this could explain the existence of multiple concepts. On the other hand, de-identification is a more clear-cut and widely harmonised theme because it is defined in a precise way via Health Insurance Portability and Accountability Act (HIPAA).[96]

The themes of anonymisation and de-identification appear to be gradually evolving, for example, older records considered anonymisation and de-identification as equivalent, while newer studies consider de-identification as a mechanical process to remove the identifiers, whereas anonymisation is the next step to prepare data for sharing (via data manipulation and privacy models). In general, most authors adhere to the narrative of further anonymising (via data manipulation and privacy models) the dataset after key variables have been removed, regardless of their previous definition of anonymisation and de-identification. Anonymisation is as a process to balance the minimisation of the probability of re-identification versus the utility of a clinical trial dataset, (e.g. too much anonymisation could render the data unusable).[36,44,46,50,66,69,71,86,87] Therefore, data cannot be fully anonymised in the context of clinical trials.

Also, it seems well accepted and understood among authors that some variables in a clinical trial dataset are identifiers and that they can be classified as direct (e.g. name or address)[16] and indirect (also named as quasi-identifiers (e.g. present age instead of date of birth)).[16]

Pseudonymisation of data usually occurs in the initial stages of data collection within clinical trials.[5] It also has a regional connotation, bound by the local laws and regulations. Pseudonymisation is declared to carry low risk for re-identification,[5,66] however, no authors from the included studies advocated its use in isolation for data sharing. Some authors acknowledge that pseudonymisation alone is not acceptable for data sharing, as the one-to-one correspondence with the original fully identified dataset still exists, which makes it personal information under the EU and UK General Data Protection Regulation (GDPR).[34,36,42,57,66]

Topic 2: Most common data manipulation techniques to achieve anonymisation. Data manipulation techniques can be applied according to the data holder's preference and technical capabilities and the intrinsic needs of the clinical trial dataset that is being processed.

Data manipulation techniques have multiple names, but there seems to be a progression towards a concerted set of four tools: perturbation, recalculation, recoding and suppression as presented in Table 3, with suppression, recoding and recalculation being the most talked about techniques. Authors are mostly describing via examples what is available regarding data manipulation techniques without critical judgement of the techniques, however, the majority of authors agree that data manipulation techniques are capable of reducing utility if left unexamined.

Topic 3: The introduction of privacy models. Clinical trials datasets are relatively small when compared to routinely collected data (e.g. medical records) and the implementation of a privacy model (such as differential privacy[90]) could present challenges, also privacy models could be complicated techniques. This can explain why the uptake of privacy models is modest, despite the fact they come from methodologies that have been tried and tested in big datasets[35,97] and they could be applied to clinical trials.[35,38,40,55,57] The most common privacy model mentioned is k-anonymity.[88]

Topic 4: The importance of controlled access and the tension with open access. The majority of clinical trial researchers strongly advocate for controlled access to the anonymised datasets, stemming from a concern with correct and genuine use of the anonymised data set.[87,92]

Authors recommend that the secondary researchers should have reasonable research questions and a data-sharing agreement should be put in place, which should include the use of the data for the intended purpose, the implementation of data protection procedures, the prohibition of any patient re-identification, the prohibition of sharing the data with a third party and the acknowledgement of the original authors in the secondary research output.

Regarding the actual sharing of the data, the trend is towards data access (e.g. via a safe haven) instead of data transfer, this means that secondary researchers can see and analyse the dataset but not download it. Here, the central repository plays a key role, because it would prove difficult (when it is necessary), to merge datasets that reside in separate repositories.

It is important to point out that controlled access is not required by laws or regulation, it is something that clinical trials researchers are doing, because it provides better research governance and researchers' acknowledgement that anonymised datasets are still sensitive.[87] Stripping identifiers from datasets and the use of manipulation techniques are not sufficient on its own to fully anonymise clinical trials datasets and to protect patient privacy. Understandably, researchers do not want to breach patient trust and they want to pre-emptively defend against a potential data breach and its catastrophic consequences (loss of patient trust, hefty legal fines and loss of reputation),[98] but they are generally willing to share.[99]

At the other end of the spectrum, open access is a relatively hassle-free release option once the dataset is anonymised, therefore, its existence and practicality is

acknowledged, but it is not directly endorsed by any of the included papers as the research governance is very difficult under it. The International Stroke Trial (IST) database[91,100] which is often cited as an example of a successful open access dataset by authors,[66,87] also drew criticisms from others[86] regarding some of the indirect identifiers left in the dataset. However, IST is yet to report a successful re-identification attack. The limited use of open access causes frustration among secondary researchers who are eager to get fast and easy access to datasets.[101,102]

Currently, controlled access is still one of the main cornerstones for the release of anonymised clinical trials data and many authors agree that data should only be released if a threshold of acceptable risk is achieved. There are several available methods for calculating risk, but authors of included studies did not explain sufficiently what 'acceptable' means, reasonably, this is very difficult to define as it would depend on the context surrounding the release of the anonymised clinical trials datasets and on the datasets own characteristics.

Comparison with existing literature

We identified a similar systematic review by Chevrier,[103] which included all biomedical literature in MEDLINE® between 2007 and 2017. We agreed with them about the existence of multiple interpretations for anonymisation and de-identification and they also discussed the balancing act between the re-identification risk and data manipulation. However, their focus was on electronic health records, and those datasets have different needs and their own challenges when compared with clinical trials datasets.

Strengths and limitations

Strengths to this review are that the electronic databases were searched since inception without any language restrictions and there was a thorough coverage of grey literature. The database searches were complemented by screening of publications on websites of key organisations, and by citation tracking. Despite our extensive search, there might be a lack of representation from other regions outside the United States–Canada, the EU and the United Kingdom. The literature databases used in this review are international in scope, but are published in North America and Europe, and are known to be stronger in coverage of literature from those regions, so that, an unknown quantity of global literature not indexed in those databases was not scrutinised as part of this review.

In the same way, identification of other sources was biased towards websites and funders in the United States–Canada, the EU and the United Kingdom, due to lack of time and funding.

If this review is to be updated, it is possible to only run the electronic searches to obtain a quick actualisation of the recommendations. The records obtained searching other sources have strengthened the evidence found from the electronic searches, contributing more than half of the included studies, but they did not provide brand new information and searching other sources was a manual and time-consuming process. However, it could be worthwhile to directly seek updated records extracted from the Medical Research Council,[71] European Medicines Agency,[57] US Department of Health & Human Services[61,76,84] and the Global Healthcare Data Science Community (Pharmaceutical Users Software Exchange – PhUSE).[23,24,58–60,67,69,70,72,73]

This review is exclusively gathering published recommendations/practices tailored specifically to clinical trials and it could not assess what researchers are actually using (but not reporting) for anonymising clinical trial data for sharing.

As this is a scoping review, there was no assessment of the quality of the evidence, therefore, we did not attempt to either explain how the included studies interpreted their local regulation on anonymisation, or to identify the existence of gaps in current practices from the obtained studies. The coding of the themes was a manual process and therefore subjective, however, a second reviewer sense-checked the coding and disagreements were mediated by a third reviewer which reduced the subjectivity of the findings.

Conclusion

Currently, there is a strong demand for academic researchers to share their data more readily. In clinical trials, data can be shared more widely if they are anonymised, yet, we do not have standardised recommendations on how to do this. As time goes by there seems to be an emerging natural consensus on the definitions of pseudonymisation, de-identification and anonymisation.

The data manipulation techniques currently used are still simple, with an increasing amount of authors recommending a shift towards privacy models, such as k-anonymity. There are other privacy models but they are not routinely used in clinical trials, as they could be complex, time-consuming and not practical for clinical trials datasets (which are relatively small when compared against routine health data).

It is impossible to discuss anonymisation in clinical trials datasets without considering the way in which the data is going to be accessed. Controlled access is still the keystone for the release of clinical trial data.

Finally, an increasing number of authors agree that data should only be released if a threshold of acceptable risk is achieved, but there is not a clear definition of 'acceptable' as this is a very complex parameter that

not only relies on the dataset but it is also embedded in a wider context out of scope for this review.

The studies identified during this review need to next be critically appraised to identify any gaps in the literature regarding anonymisation methods and data access approaches. Also, clear guidance on methods for quantifying the risk of re-identification need to be developed. This would allow for the creation of standardised worldwide recommendations for data sharing in clinical trials reflecting the growing consensus exhibited in the literature found during this review.

Author contributions

A.R., S.C.L. and C.J.W. conceived the idea for this work supported by S.E., M.F.D., T.J. and C.T. A.R. wrote the first draft, and all authors contributed to the article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Ethics and dissemination

This project did not collect any patient data or outcomes; therefore, it was not necessary to seek formal National Health Service Research Ethics Committee's approval. However, we applied for ethical approval from the Internal Ethics Review Board at The University of Edinburgh's Usher Institute.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: A.R. has a scholarship from the University of Edinburgh to undertake a PhD with the support from the Asthma UK Centre for Applied Research (AUKCAR) (AUKCAR-17-01a). Neither funder (University of Edinburgh) nor sponsor (AUKCAR) contributed to protocol development. C.J.W. is supported in this work by NHS Lothian via the Edinburgh Clinical Trials Unit. S.C.L. and C.T. are supported in this work by their employment at the Edinburgh Clinical Trials Unit. S.E. is supported in this work by her employment at the Pragmatic Clinical Trials Unit. M.F.D. is supported in this work by their employment at the University of Edinburgh. T.J. is supported by Asthma UK as part of the Asthma UK Centre for Applied Research (grant nos AUK-AC-2012-01 and AUK-AC-2018-01). A.M. is an independent researcher.

Protocol registration

The protocol for this scoping review was not registered with the International Prospective Register of Systematic Reviews (PROSPERO) as the proposed systematic review did not meet the PROSPERO inclusion criteria. However, the final protocol dated 11 January 2019 is attached as Appendix 1 in the supplemental materials.

ORCID iDs

Aryelly Rodriguez https://orcid.org/0000-0002-1352-3922
Tracy Jackson https://orcid.org/0000-0002-6188-3607
Christopher J Weir https://orcid.org/0000-0002-6494-4903

Supplemental material

Supplemental material for this article is available online.

References

1. Chan A-W, Song F, Vickers A, et al. Increasing value and reducing waste: addressing inaccessible research. *Lancet* 2014; 383: 257–266.
2. Hrynaszkiewicz I and Altman DG. Towards agreement on best practice for publishing raw clinical trial data. *Trials* 2009; 10: 1–5.
3. Song F, Hooper L and Loke Y. Publication bias: what is it? How do we measure it? How do we avoid it? *Open Access J Clin Trials* 2013; 2013: 71–81.
4. Berlin JA, Morris S, Rockhold F, et al. Bumps and bridges on the road to responsible sharing of clinical trial data. *Clin Trials* 2014; 11(1): 7–12.
5. Ohmann C, Banzi R, Canham S, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open* 2017; 7: e018647.
6. El Emam K and Arbuckle L. *Anonymizing health data: case studies and methods to get you started.* Sebastopol, CA: O'Reilly Media, Inc., 2013.
7. Hrynaszkiewicz I, Norton ML, Vickers AJ, et al. Preparing raw clinical data for publication trials. *Trials* 2010; 11: 9.
8. Information Commissioner's Office (ICO). Anonymisation code of practice, 2012, https://ico.org.uk/media/1061/anonymisation-code.pdf
9. Keerie C, Tuck C, Milne G, et al. Data sharing in clinical trials – practical guidance on anonymising trial datasets. *Trials* 2018; 19: 25.
10. Tudur Smith C, Hopkins C, Sydes M, et al. Good practice principles for sharing individual participant data from publicly funded clinical trials, 2015, https://www.methodologyhubs.mrc.ac.uk/files/7114/3682/3831/Datasharingguidance2015.pdf
11. Google. Google Scholar. https://scholar.google.com/
12. The Joanna Briggs Institute. *Joanna Briggs Institute Reviewers' manual: 2015 edition / supplement. Methodology for JBI scoping reviews*. Adelaide, SA, Australia: The Joanna Briggs Institute, 2015.
13. Peters MD, Godfrey CM, Khalil H, et al. Guidance for conducting systematic scoping reviews. *JBI Evid Implement* 2015; 13: 141–146.
14. Tricco AC, Lillie E, Zarin W, et al. Preferred reporting items for systematic review and meta-analysis (PRISMA) extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018; 169: 467–473.
15. Tudur Smith C, Hopkins C, Sydes MR, et al. How should individual participant data (IPD) from publicly funded clinical trials be shared? *BMC Med* 2015; 13: 298.
16. Hrynaszkiewicz I, Norton ML, Vickers AJ, et al. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials* 2010; 11: 9.

17. Health Research Authority. Funding, 2019, https://www.hra.nhs.uk/planning-and-improving-research/research-planning/funding/
18. Wellcome. Clinical trials policy – Grant funding| Wellcome, 2019, https://wellcome.org/grant-funding/guidance/clinical-trials-policy
19. Wikipedia. List of wealthiest charitable foundations. *Wikipedia*, 2019, https://en.wikipedia.org/wiki/List_of_wealthiest_ charitable_foundations
20. May M. Top 100 UK charities ranked for brand value in new league table. *UK Fundraising*, 27 November 2018, https://fundraising.co.uk/2018/11/27/top-100-uk-charities-ranked-brand-value-new-league-table/
21. UKCRC. Registered CTUs – UKCRC, 2019, https://ukcrc-ctu.org.uk/registered-ctus/.
22. Meeting abstracts from the 5th International Clinical Trials Methodology Conference (ICTMC 2019). *Trials* 2019; 20: 579.
23. PhUSE. *PhUSE DeID Standard – SDTM 3.2 – appendix 1 – date offsetting – v1.91[2]*. Kent: PhUSE, 2015.
24. PhUSE. *PhUSE Data De-Identification Standard for SDTM 3.2 – appendix 2-low frequencies-v10-19387*. Kent: PhUSE, 2015.
25. The EndNote Team. *EndNote. EndNote X8 ed.* Philadelphia, PA: Clarivate Analytics, 2016.
26. Veritas Health Innovation. *Covidence systematic review software*. Melbourne, VIC, Australia: Veritas Health Innovation.
27. Microsoft Corporation. Microsoft Excel. 2016, https://office.microsoft.com/excel
28. Higgins JPT and Green S. *Cochrane handbook for systematic reviews of interventions*. London: The Cochrane Collaboration, 2011.
29. QSR International Pty Ltd. NVivo (Version 11). 2015, https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home
30. Thomas J and Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol* 2008; 8: 1–10.
31. Gibbs GR. Thematic coding and categorizing. *Anal Qual Data* 2007; 703: 38–56.
32. Parcell ES and Baker B. Narrative analysis. In: Allen M (ed.) *The Sage encyclopedia of communication research methods*, vol. 3. Thousand Oaks, CA: SAGE, 2017, pp. 1069–1072.
33. Atzor S, Sorof J, Kelman A, et al. Clinical trial data sharing: from principles to practical implementation – an industry model. *Regul Rapp* 2014; 11: 4–7.
34. Demotes-Mainard J, Cornu C, Guerin A, et al. How the new European data protection regulation affects clinical research and recommendations? *Therapie* 2019; 74: 31–42.
35. El Emam K and Dankar FK. Protecting privacy using k-anonymity. *J Am Med Inform Assoc* 2008; 15(5): 627–637.
36. El Emam K, Rodgers S and Malin B. Anonymising and sharing individual patient data. *BMJ* 2015; 350: h1139.
37. Lee J, Jung J, Park P, et al. Design of a human-centric de-identification framework for utilizing various clinical research data. *Hum-Centric Comput Inf Sci* 2018; 8: 1–12.
38. Malin B, Karp D and Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J Investig Med* 2010; 58(1): 11–18.
39. Morse RE, Nadkarni P, Schoenfeld DA, et al. Web-browser encryption of personal health information. *BMC Med Inf Decis Mak* 2011; 11: 70.
40. Nasseh D, Engel J, Mansmann U, et al. Matching study to registry data: maintaining data privacy in a study on family based colorectal cancer. *Stud Health Technol Inform* 2014; 205: 808–812.
41. Nitzlnader M and Schreier G. Patient identity management for secondary use of biomedical research data in a distributed computing environment. *Stud Health Technol Inform* 2014; 198: 211–218.
42. Noumeir R, Lemay A and Lina JM. Pseudonymization of radiology data for research purposes. *J Digit Imaging* 2007; 20(3): 284–295.
43. Schell SR. Creation of clinical research databases in the 21st century: a practical algorithm for HIPAA Compliance. *Surg Infect* 2006; 7(1): 37–44.
44. Sudlow R, Branson J, Friede T, et al. EFSPI/PSI working group on data sharing: accessing and working with pharmaceutical clinical trial patient level datasets – a primer for academic researchers. *BMC Med Res Methodol* 2016; 16: 73.
45. Tuck C, Lewis S, Milne G, et al. Data sharing in clinical trials – practical guidance on anonymising trial datasets – oral presentation. *Trials* 2015; 16. https://doi.org/10.1186/1745-6215-16-S2-O66
46. Tucker K, Branson J, Dilleen M, et al. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med Res Methodol* 2016; 16(Suppl. 1): 77.
47. Tudur Smith C, Hopkins C, Sydes M, et al. Good practice principles for sharing individual participant data from publicly funded clinical trials. *Trials* 2015; 16. https://doi.org/10.1186/1745-6215-16-S2-O1
48. Tudur Smith C, Nevitt S, Appelbe D, et al. Resource implications of preparing individual participant data from a clinical trial to share with external researchers. *Trials* 2017; 18: 319.
49. Wallace SE, Gaye A, Shoush O, et al. Protecting personal data in epidemiological research: DataSHIELD and UK law. *Public Health Genomics* 2014; 17(3): 149–157.
50. Asthma UK Centre for Applied Research. ASTHMA UK policy data sharing – introduction to sharing individual participant data, version 2. Circa, 2015, p. 2.
51. Australian National Medical Research Data Storage Facility. Anonymisation. Circa, 2016, https://researchdata.edu.au/meddataeduau/632288
52. Clinical Study Data Request (CSDR). CSDR – anonymisation of clinical trial datasets. Circa, 2015, https://www.clinicalstudydatarequest.com/Default.aspx
53. Clinical Study Data Request (CSDR) and Eisai. CSDR – anonymisation of clinical trial datasets – Eisai circa, 2015, https://www.clinicalstudydatarequest.com/Study-Sponsors/Study-Sponsors-Eisai.aspx
54. Clinical Study Data Request (CSDR) EL. CSDR – anonymisation of clinical trial datasets – Eli Lilly and Company. *Circa*, 2015, https://www.clinicalstudydatarequest.com/Documents/Anonymisation_clinicaltrialdata_Lilly_update.pdf
55. Ebner H, Hayn D, Falgenhauer M, et al. Piloting the European unified patient identity management (EUPID) concept to facilitate secondary use of neuroblastoma data from Clinical Trials and Biobanking. In: Schreier G,

Ammenwerth E, Hörbst A, et al. (eds) *Health informatics meets eHealth: predictive modeling in healthcare–from prediction to prevention proceedings of the 10th eHealth2016 conference*. Amsterdam: IOS Press, 2016, pp. 31–38.
56. El Emam K and Malin B. Concepts and methods for de-identifying clinical trial data. Paper commissioned by the Committee on Strategies for Responsible Sharing of Clinical Trial Data, 2014.
57. European Medicines Agency (EMA). *Data anonymisation – a key enabler for clinical data sharing*. Workshop Report No. EMA/796532/2018, 4 December 2018. London: European Medicines Agency.
58. Ferran J-M. PhUSE – De-Identification Standards for CDISC data models – PhUSE, Data Transparency Working Group Lead. In: *4th international clinical trials methodology conference (ICTMC)*, Liverpool, 7–10 May 2017.
59. Ferran J-M, El Emam K, Nolan S, et al. *PhUSE De-Identification Working Group: providing De-Identification Standards to CDISC data models*. Kent: PhUSE, 2015.
60. Ferran J-M and Lanoue J. PhUSE De-Identification Working Group: providing De-Identification Standards to CDISC data models – DS10, 2015, https://www.pharmasug.org/proceedings/2015/DS/PharmaSUG-2015-DS10.pdf.
61. Food Drug Administration. HHS – availability of masked and de-identified non-summary safety and efficacy data; request for comments. *Federal Register*, 2013, https://www.federalregister.gov/articles/2013/06/04/2013-13083/availability-of-masked-and-de-identified-non-summary-safety-and-efficacy-data-request-for-comments
62. Hollis S, Fletcher C, Lynn F, et al. Best practice for analysis of shared clinical trial data. *BMC Med Res Methodol* 2016; 16(Suppl. 1): 76.
63. Hughes S, Wells K, McSorley P, et al. Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach. *Pharm Stat* 2014; 13(3): 179–183.
64. Huser V and Shmueli-Blumberg D. Data sharing platforms for de-identified data from human clinical trials. *Clin Trials* 2018; 15(4): 413–423.
65. International Pharmaceutical Privacy Consortium. IPPC white paper on anonymisation of clinical trial datasets, 2014, https://6a908337-3075-4032-a3f9-cee264a142f8.filesusr.com/ugd/932589_48d9c33238994cdfa5ee4273a29fe444.pdf
66. IOM (Institute of Medicine). *Sharing clinical trial data: maximizing benefits, minimizing risk*. Washington, DC: National Academies Press, 2015.
67. Iversen JM. *PhUSE – Data De-Identification made simple*. Ballerup, Denmark: PHUSE – LEO Pharma A/S, 2016.
68. Jonas S, Siewert S and Spreckelsen C. Privacy-preserving record grouping and consent management based on a public-private key signature scheme: theoretical analysis and feasibility study. *J Med Internet Res* 2019; 21: e12300.
69. Kniola L, Hughes A, Paczewska-Sosnowska A, et al. *PhUSE – data anonymisation and risk assessment automation*. Kent: PhUSE, 2020, p. 10.
70. Lyathakula S. PhUSE – data anonymization providing clinical trial data to outside researchers. In: NOVARTIS (ed.) *PhUSE Single Day Event (SDE)*. Mumbai, India, 2015.
71. Medical Research Council (MRC). GDPR guidance note 5: identifiability, anonymisation and pseudonymisation, 2019, https://www.ukri.org/wp-content/uploads/2021/11/MRC-291121-GDPR-Identifiability-Anonymisation-Pseudonymisation.pdf
72. Meeh S. PhUSE Data De-Identification Standard for CDSIC SDTM IG 3.2, and EMA Policy 0070. Integrated Data Analytics and Reporting Janssen, 2016, https://slideplayer.com/slide/17911736/
73. Meeh S. PhUSE Data De-Identification Standard for CDSIC ADaM 2.1 IG 1.0, and updates for SDTM IG 3.2, 2017, https://slideplayer.com/slide/11742761/
74. Miller JD. Sharing clinical research data in the United States under the health insurance portability and accountability act and the privacy rule. *Trials* 2010; 11: 112.
75. National Institutes of Health (NIH). NIH data sharing policy and implementation guidance, 2003, https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
76. National Institutes of Health (NIH). HHS – clinical research and the HIPAA privacy rule, 2004, http://privacyruleandresearch.nih.gov/clin_research.asp
77. Nelson GS. Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification. In: SAS (ed.) *SAS GLOBAL FORUM proceedings 2015*, 2015, pp. 1–23, https://www.pharmasug.org/proceedings/ 2016/IB/PharmaSUG-2016-IB06.pdf
78. Olesen S. Publishing and sharing sensitive data. Australian National Data Service, 2011, https://www.ands.org.au/guides/sensitivedata
79. Pfizer. Clinical trial data access – policy document 01312014, 2014.
80. Shostak J. De-identification of clinical trials data demystified. SAS Users Group, 2006, https://www.lexjansen.com/pharmasug/2006/PublicHealthResearch/PR02.pdf
81. The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation. *Accessing health and health-related data in Canada*. Ottawa, ON, Canada: Council of Canadian Academies, 2015.
82. TransCelerate BioPharma Inc. TransCelerate – data de-identification and anonymization of individual patient data in clinical studies. TransCelerate – Clinical Data Transparency Initiative, 2016, https://www.transcelerate biopharmainc.com/initiatives/clinical-data-transparency/
83. TransCelerate BioPharma Inc. TransCelerate – anonymization of individual patient data in clinical studies – a model approach. De-identification, TransCelerate-Data, 2015, https://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-V2.0.pdf
84. U.S. Department of Health & Human Services (HHS). *HHS – guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. Washington, DC: U.S. Department of Health and Human Services, 2012, p. 26, https://www hhsgov/hipaa/for-professionals/privacy/special-topics/de-identification/indexhtml

85. Walker N. All or nothing: the false promise of anonymity. *bioRxiv* 2016: 084921, https://www.biorxiv.org/content/biorxiv/early/ 2016/11/03/084921.full.pdf
86. Walker N. All or nothing: the false promise of anonymity. *Data Sci J* 2017; 16: 24.
87. Tudur Smith C, Hopkins C, Sydes MR, et al. *Good practice principles for sharing individual participant data from publicly funded clinical trials*, version 1. Medical Research Council; Hubs for Trials Methodology Research, 2015, https://www.methodologyhubs.mrc.ac.uk/files/7114/3682/3831/Datasharingguidance2015.pdf
88. Sweeney L. K-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst* 2002; 10: 557–570.
89. Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity; privacy beyond k-anonymity. *Acm T Knowl Discov D* 2007; 1: 3–es.
90. Dwork C. Differential privacy: a survey of results, 2008, pp. 1–19, https://web.cs.ucdavis.edu/~franklin/ecs289/2010/dwork_2008.pdf
91. Clinical Trials Unit London School of Hygiene & Tropical. freeBIRD (Bank of injury and Emergency Research Data) 2011, https://freebird.lshtm.ac.uk/home/
92. Clinical Study Data Request (CSDR). Clinical study data request. https://www.clinicalstudydatarequest.com/
93. Project Data Sphere. An independent initiative of the CEO Roundtable on Cancer. https://www.projectdatasphere.org/
94. The Yale University. Yale University Open Data Access (YODA) Project, https://yoda.yale.edu/
95. Vivli Center for Global Clinical Research Data. A global clinical research data sharing platform, https://vivli.org/home-mar2020/
96. Health UDo and Services H. *Protecting personal health information in research: understanding the HIPAA privacy rule*. Washington, DC: Author, 2003.
97. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertain Fuzziness Knowl Based Syst* 2002; 10: 571–588.
98. Information Commissioner's Office (ICO). Anonymisation: managing data protection risk code of practice, 2012, https://ico.org.uk/media/1061/anonymisation-code.pdf
99. Rathi V, Dzara K, Gross CP, et al. Sharing of clinical trial data among trialists: a cross sectional survey. *BMJ* 2012; 345: e7570.
100. Roberts I, Shakur H, Coats T, et al. The CRASH-2 trial: a randomised controlled trial and economic evaluation of the effects of tranexamic acid on death, vascular occlusive events and transfusion requirement in bleeding trauma patients – appendix 4 Free Bank of Injury and emergency Research Data – freeBIRD. *Health Technol Assess* 2013; 17: 1.
101. Loder E. Data sharing: making good on promises. *BMJ* 2018; 360: k710.
102. Dunn AG, Day RO, Mandl KD, et al. Learning from hackers: open-source clinical trials. *Sci Transl Med* 2012; 4: 132–135.
103. Chevrier R, Foufi V, Gaudet-Blavignac C, et al. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *J Med Internet Res* 2019; 21: e13484.

3.3. Conclusions

The preliminary results, derived solely from the execution of the electronic literature database searches were successfully presented as an oral presentation at the 5th International Clinical Trials Methodology Conference (ICTMC) in October 2019 in Brighton, UK. The subsequent addition of the grey literature did not alter the emerging themes found through the studies from the database searches.

The fully executed scoping review has three core findings. The first is that there are no standardised recommendations on how to anonymise clinical trial datasets for sharing. This aligns with the observations of Sariyar et al. (2015) (Sariyar, Schluender et al. 2015), who determined that anonymisation is a regional definition, making it difficult to provide a precise definition. Chevrier et al. (2019) (Chevrier, Foufi et al. 2019) also observed a similar issue and recommended the development of clear definitions and guidance in their respective fields—Sariyar et al. (2015) in human bio samples and Chevrier et al. (2019) in biomedical literature.

The second and third core findings of the scoping review are that a consensus is developing on definitions and techniques used to implement anonymisation, and that anonymisation techniques alone are insufficient to protect patient privacy; they need to be paired with controlled access.

Since the publication of the scoping review, articles by El Emam and Abdallah (2015) (El Emam and Abdallah 2015) and Handelsman (2015) (Handelsman 2015) have been brought to my attention during a webinar on 17th April 2024, titled “The Costs of Anonymisation” (Pilgram, Meurers et al. 2024). Upon retrospectively checking these articles, I found they would have met the inclusion criteria of the scoping review. These articles are not indexed in any literature database, as verified by Ms. Marshall Dozier (UoE leader of the academic support librarians), which explains their

omission and makes their findability an issue. However, the absence of the article by El Emam and Abdallah (El Emam and Abdallah 2015) did not change the conclusions of the scoping review because it is very similar to the already included El Emam et al. (2015) (El Emam, Rodgers et al. 2015) article. While El Emam and Abdallah's findings align with ours, our scoping review uses a systematic approach, bringing together the collective knowledge of the community. Handelsman's work discusses similar topics to our scoping review, framing them in the practical context of using SAS software (SAS Institute Inc 2013) to deliver de-identification.

I also found another eligible article by Gudi et al. (2022) (Gudi, Kamath et al. 2022), which was published in May 2022. This article would have been eligible but was left out because it was published outside our search period, which ended in February 2021. Nonetheless, Gudi et al. (2022)(Gudi, Kamath et al. 2022) reached similar conclusions to ours, albeit using a different search scope targeted towards data sharing policy or guidance in clinical trials.

At the time of writing this thesis (August 2024), the article associated with this chapter has been formally cited 16 times (according to Google Scholar(Google)) and has been viewed and downloaded 4,200+ times (Clinical Trials, a SAGE journal, 2022 (Clinical Trials a SAGE journal 2022)). It has also served as supporting evidence for "Qualitative data sharing practices in clinical trials in the UK and Ireland: towards the production of good practice guidance" (McCarthy et al., 2023) (McCarthy, Gillies et al. 2023) to extend the interpretation of anonymisation, de-identification, and data sharing practices to qualitative data.

Having the published recommendations in one place through this scoping review helped us understand how publicly available anonymised datasets from clinical trials might have been put together. In the next chapter, we will explore how the

implementation of these guidelines and recommendations is manifesting in the process of data sharing and how participant privacy is protected through anonymisation and de-identification, because the scoping review revealed that while some authors accept that anonymised datasets can only be release if an acceptable level of risk is achieved, very little guidance exist on how this re-identification risk should be assessed.

Chapter 4 Exploration and assessment re-identification risks scores of anonymised clinical trials datasets shared for secondary research

4.1 Introduction

In Chapter 3, I collected published recommendations and guidance on how to anonymise clinical trial datasets(Rodriguez, Tuck et al. 2022). A key aspect of this scoping review was that 35% of the collected sources highlighted the importance of assessing the risk of the anonymised datasets before deciding on their release. However, only three sources (7%) provided methods for estimating this risk.

With an understanding of how anonymised datasets could be compiled, the advice to evaluate the re-identification risk, and awareness of the increasing public availability of these datasets due to rising pressures and incentives(Huser and Shmueli-Blumberg 2018), I had the preliminary knowledge needed to address the second and third objectives of this PhD: "To investigate whether individual participants could potentially be at risk of being re-identified from a range of datasets that have been anonymised and made available for sharing" and "To identify factors that could increase the risk of re-identification of an anonymised clinical trial dataset."

To address these questions, I conducted exploratory(Babbie 2020, Swedberg 2020, George 2021) quantitative research(Vogt 2011, Sheard 2018, Bhandari 2022) on anonymised clinical trial datasets shared for secondary research purposes(Saunders, Lewis et al. 2009, Goodwin 2012). This research aimed to explore the datasets' characteristics and assess the feasibility of implementing re-identification risk methods. Well-established methodologies exist for exploring the re-identification risk of anonymised datasets in broader contexts(Information Commissioner's Office (ICO) 2012, O'Keefe, Otorepec et al. 2017, Office of the Information Commissioner Queensland 2020) (Elliot 2016, Elliot, Mackey et al. 2020), as

explained in Chapter 2. This methodology has been particularly studied for the release of electronic health records in the US and Canada(El Emam and Fineberg 2009, Skinner 2009, El Emam, Buckeridge et al. 2011, El Emam, Arbuckle et al. 2012, Simon, Shortreed et al. 2019), but it is not commonly recommended for clinical trial datasets, as seen in our scoping review(Rodriguez, Tuck et al. 2022).

Therefore, I developed the protocol titled "What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol" in Appendix 3. I used this protocol to apply for ethical research approval on 3rd December 2022, using the "Usher Research Ethics Group 'Triage' Tool" in Appendix 3. This tool helped identify the appropriate form for the ethics application, pointing to the "Fully Anonymous Secondary Data Analysis (FASDA)" form (Appendix 3), as I intended to perform secondary data analysis on data extracts containing 'fully anonymous' information. This led down a self-certification path due to the "absence of reasonably foreseeable ethics risks."

In the protocol, I proposed to take a sample of datasets and apply the re-identification risk calculations described by El Emam (2013)(El Emam 2013). As this is an exploratory study with limited resources, the sample size (number of datasets) was deliberately small and arbitrary. Therefore, if a single data repository had five or fewer datasets meeting the inclusion criteria, I requested all datasets from that repository. If a repository had more than five eligible datasets, a random selection of five datasets was drawn from the eligible datasets.

After completing all protocol activities, I compiled a manuscript for the publication of the results, as I believe this part of my research will be highly beneficial to the wider clinical trials community. I meticulously addressed all comments from my co-authors and data holder/owners and submitted the manuscript to the peer-reviewed Clinical Trials SAGE Journal on 23rd July 2024, and it is currently under review. Our submitted manuscript,

titled “Evaluating Re-Identification Risk Scores in Publicly Available Clinical Trial Datasets: Insights and Implications,” is presented in the next section of this chapter. As shown in other chapters in this thesis I am planning to support this manuscript all the way until publication.

All additional materials associated with this submitted article are included in Appendix 3.

4.2 Submitted manuscript

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Rodriguez A[1], Williams LJ[1], Lewis SC[1], Sinclair P[1], Eldridge S[2], Jackson T[3], Weir CJ[1]

[1]Edinburgh Clinical Trials Unit, Usher Institute, the University of Edinburgh

[2]Pragmatic Clinical Trials Unit, Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London

[3]Asthma UK Centre for Applied Research, Usher Institute, the University of Edinburgh

Correspondence:

Ms Aryelly Rodriguez

Edinburgh Clinical Trials Unit (ECTU)

Usher Institute, University of Edinburgh

5-7 Little France Road, Edinburgh, EH16 4UX

Emails: aryelly.rodriguez@ed.ac.uk, Linda.Williams@ed.ac.uk, steff.lewis@ed.ac.uk, christopher.weir@ed.ac.uk, Pamela.Sinclair@ed.ac.uk, Tracy.Jackson@ed.ac.uk, s.eldridge@qmul.ac.uk

https://journals.sagepub.com/author-instructions/ctj

Sage Clinical Trials - original research papers (maximum 4,000 words excepting abstract, references, tables and figures; maximum 6 tables or figures)

Manuscript approx. word count: 3988

Key Words: Clinical Trials | Data Anonymisation | Re-identification | De-identification | Datasets | Risk scores | Data Sharing

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Abstract

Background: The motivations to share anonymised datasets from clinical trials within the scientific community are increasing. Many anonymised datasets are now publicly available for secondary research. However, it is uncertain whether they pose a privacy risk to the involved participants.

Methods: We located a broad sample of publicly available, de-identified/anonymised randomised clinical trials datasets from human participants and contacted their owners to request access, following their local procedures. We classified personal data within these datasets, including unique direct identifiers such as dates of birth, and other personal data that, on their own, does not identify an individual but may do so when combined with each other, such as sex, age and race (indirect identifiers). Combining indirect identifiers forms strata, and the re-identification risk score equations evaluate membership in these strata in three ways. First, by measuring the proportions of participants in strata above predetermined risk threshold levels (Ra). Second, by locating the smallest stratum (Rb). Third, by estimating the average membership across all strata in a dataset (Rc). The risk scores range from 0 (lowest risk) to 1 (highest risk); they do not aim to re-identify individuals in the datasets and are used for routinely collected health records. If a dataset contained a direct identifier, it automatically scored 1 in all metrics. Conversely, if a dataset contained no direct or indirect identifiers, or only one indirect identifier, it automatically scored 0 in all metrics. Finally, we explored which characteristics of the datasets were associated with the risk scores and compared the risk scores and their usability.

Results: Seventy datasets from 14 data sources were analysed. Thirty-one datasets were shared with minimal restrictions (open access), while 39 were shared with varying levels of restrictions before access was granted (controlled access). Datasets had, on average, four identifiers and mean risk scores ranging from 0.47 to 0.91. The most common pieces of information present in the datasets that, when combined, may indirectly identify a participant were sex (80%) and age (72.9%).

Conclusions: This study confirms that clinical trial datasets are rich in personal details and that using re-identification risk scores as a measure of this richness is feasible. These scores could inform the anonymisation process of clinical trials datasets to release them for secondary research. We propose a strategy for employing the scores in the decision-making process for releasing clinical trials datasets.

Abstract word count: 389 (Maximum 425)

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Background

There is a strong drive, particularly from publishers and funders, to encourage the release of anonymised trial data sets[1]. New grant applications with funding from the National Institutes of Health[2], Cancer Research UK[3] and the UK Medical Research Council[4] must contain a concrete data-sharing plan. The International Committee of Medical Journal Editors (ICMJE) requires that all clinical trials starting enrolling participants on or after January 1, 2019, to have a data sharing plan in their registration[5]. Also, the ICMJE encourages editors to prioritise publishing work from authors who have shared their data[6]. Data-sharing has become essential in clinical trials for disseminating current research, enhancing knowledge through meta-analysis, enabling new investigations on existing datasets, and maximising the efforts invested in data collection[7, 8]. Many anonymised datasets are currently available for secondary research via clinical data repositories[9] or directly from researchers, using either open or controlled access models. Controlled access requires some form of approval before obtaining datasets, whereas open access imposes minimal restrictions[10]. Anonymisation of data is complex[10], and complete anonymisation could result in a loss of detail necessary for appropriate analysis. Therefore, balancing the reduction of re-identification risk while retaining sufficient detail for valid research is crucial. The lack of a gold-standard anonymisation method within the clinical trial community[11] complicates this further, leading to uncertainty about the privacy risks posed by available datasets[12, 13] to the involved participants.

We evaluated publicly available datasets by calculating their re-identification risk scores using El-Emam's methods[14]. These scores numerically estimate the potential of re-identifying individuals from anonymised data, helping to assess the effectiveness of anonymisation techniques in preserving privacy. They are used for routinely collected health records[15, 16], and theoretically indicate higher re-identification risks with higher scores, without aiming to re-identify individuals.

Importance of the Study

To our knowledge, no studies have directly used the proposed methods of calculating re-identification risk scores across various publicly available clinical trial datasets.

Objectives

To calculate and describe the re-identification risk scores of publicly available clinical trials' datasets. To investigate which dataset characteristics are associated with increased or decreased risk scores. To compare the risk scores and assess their usability.

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Methods

A full protocol (Appendix 1)[17] was finalised on December 1, 2020.

We collected a broad sample of publicly available, de-identified/anonymised clinical trials datasets to estimate their re-identification risk scores, using three equations designed for this purpose [18] .

Datasets sources

We identified 18 data sources based on previous research[19], via web searches and word of mouth. We included 16 repositories allowing open or controlled access to datasets. Datasets were requested from repositories between May 7, 2021, and September 23, 2022. We also searched RCTs published in two journals with strong data-sharing policies (PLOS One and BMJ) from January 2013 for BMJ and March 2014 for PLOS One (when their data-sharing policies were introduced) to 30 April 2022 (details in Appendix 2).

Types of datasets

Our inclusion criteria consisted of datasets from RCTs on human participants, deemed anonymised and/or de-identified by data holders, and suitable for secondary research. We excluded datasets described as containing identifiable information protected solely by controlled access or data-sharing arrangements. We limited our selection to studies with materials available in English or Spanish due to the language skills within the writing team.

Data collection and analysis

Selection and request of datasets

One investigator (AR) searched the sources (repositories or journals) and screened titles and descriptions of datasets to determine eligibility and identify duplicates. For sources with five or fewer eligible datasets, all datasets were requested. For sources with more than five eligible datasets, five were selected at random using SAS software[20] by assigning a random number to each dataset, ordering them based on these random numbers and requesting the first five. If a dataset was unavailable, it was replaced with the next on the ordered list. This exploratory sample ensured a fair representation from all sources and maximised the information we could obtain given our limited resources.

Due to various factors, we revised this sampling strategy for some sources:

- Project Data Sphere[21]: has two levels of controlled access (researcher vetoing and researcher vetoing plus data sharing agreements (DSA)). Initially, five datasets were requested. The signed DSA

covered four datasets, while one did not require a DSA. Additional datasets were accessible without a DSA, so we randomly chose another four, totalling nine datasets.

- Large Repositories: Requesting more than five datasets was more efficient as some owners were unwilling to share certain datasets or the datasets did not exist. Oversampling ensured robust estimates and mitigated the inflexibility of DSAs tailored to specific datasets. For instance, at "The Yale University Open Data Access (YODA) Project" [22], we signed a DSA for five datasets, but one study only had the clinical study report without individual participant data. Replacing this dataset would have triggered a new DSA, so YODA is represented by only four datasets.
- BMJ and PLOS One: Studies offering controlled access via DSA were not pursued due to bottlenecks in our contracts department, and we had already acquired 39 controlled access datasets.

Once a dataset was identified as potentially meeting the inclusion criteria, it was downloaded (if open access) or access was applied for following the data owners' procedures. The time from request to data access approval was recorded. Some datasets could only be analysed remotely in trusted research environments (TREs).

Data extraction and management

The selected datasets were retrieved and transferred to a secure and password-protected electronic storage area at the University of Edinburgh[23] [24] [25]. For datasets held in TREs, AR transferred the re-identification risk scores calculation analysis code to those environments and extracted the relevant output.

Datasets were provided in multiple software formats and thoroughly explored to ensure no discrepancies among formats and to check for additional data. Whenever possible, the datasets were compared with their corresponding data dictionary, case report form, and/or protocol to verify if they covered all collected data or were partial datasets.

All available metadata, from obtained datasets, was recorded on an MS Excel[26] spreadsheet (attributes collected in Appendix 3).

Data synthesis and re-identification risk calculation

We extracted the number of direct and/or indirect identifiers in the datasets as described by Hrynaszkiewicz et al.[27], except for "small denominators population size of <100" and "very small numerators—event counts of <3" which were already considered within the risk score calculations. The data extraction process involved visual inspection by AR and verification by a second reviewer (SCL, LJW or CJW).

El Emam's methodology[14], which combines all indirect identifiers to create strata within a dataset, was utilised. For example, combining age, race and sex generate groups such as: 'four 18-year-old white males' or 'one 79-year-old African American female'. Adding more identifiers further divides data into smaller strata, known as granularity. The goal is to minimise the number of strata while maintaining data utility. Three re-identification risk scores were calculated[14]:

1. Risk a (Ra): Measures the proportion of participants in a dataset who belong to strata with a re-identification probability higher than seven predefined thresholds (0.01, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5)[14] (These thresholds were chosen to represent a wide variety of risks). For instance, a Ra of 0.5 at threshold 0.1 indicates that 50% of the participants in the dataset are in strata with 10 or fewer participants.
2. Risk b (Rb): Identifies the stratum with the smallest membership (regarding all indirect identifiers), representing the worst-case scenario. For example, a Rb of 0.33 indicates that there is at least one stratum with 1 in 3 chances of being re-identified.
3. Risk c (Rc): Represents the average risk score across the whole strata of the dataset, using all indirect identifiers. A dataset with two strata of 5 and 10 participants would have a Rc of 0.15, calculated as the average of 1/5 and 1/10.

Each risk score ranges from 0 to 1 and was estimated under the prosecutor and journalist scenario[28]. For the prosecutor scenario, we sought to detect uniqueness within strata in each dataset. In the journalist scenario, we used synthetic datasets, scaled to at least 15 times the size of the anonymised datasets, as identification sources for matching against the anonymised datasets. The synthetic datasets were customised to include corresponding indirect identifiers using the algorithm by Bogle and Erickson[29]. Appendix 4 shows a worked example for calculating Ra, Rb and Rc.

Several assumptions guided the risk score calculations:

- Calculations required at least two indirect identifiers; datasets with fewer were automatically scored 0 [30] for all re-identification risks in both scenarios.
- Datasets containing at least one direct identifier were automatically scored 1 [30] for all re-identification risks in both scenarios.
- No recoding or further manipulation of datasets was allowed, except for necessary steps to prepare data for re-identification risks calculation.

Metadata from all included datasets and re-identification risk scores were summarised using descriptive statistics. There were no attempts to re-identify or contact individual participants. This was an exploratory study, so no formal statistical inference was set a priori.

SAS software[20] was used for all analyses, including the calculations for risk scores, except for datasets at YODA's[22] TRE, where STATA software[31] was used because of SAS unavailability in that environment.

Results

The first dataset was received on the May 7, 2021, and the last on the Apr 26, 2023. Of the 18 identified data sources, three were excluded: two did not contain relevant RCTs datasets, and one no longer existed when the study protocol was executed. Consequently, 15 data sources were visited, 14,896 datasets preselected, and a sample of 86 datasets requested. All data sources offered the data free of charge. The median number of requested datasets per data source was 6. We obtained 76 out of 86 (88.4%) requested datasets, faced 9 (10.5%) rejections and received one (1.2%) duplicate. The reasons for rejection were: five denied access as our proposal was not considered a valid reason for data sharing, three because the datasets did not exist and one where the data owners could not be reached. The median number of obtained datasets per repository was five. (Appendix 5).

We analysed 70 out of 76 (92.1%) datasets (representing 14 data sources) and excluded six datasets (7.9%), because four were not from RCTs and two were summarised cluster data instead of individual participant data (flowchart in Appendix 6). Appendix 5 provides a list of the 70 included studies.

Included datasets' characteristics

Table 1 summarises the characteristics of the included datasets. Of the 70 datasets, 39 (55.7%) were shared with varying levels of controlled access, while 31 (44.3%) were provided with minimal restrictions (open access). On average, it took 270 days to obtain controlled access datasets and 8 days for open access datasets. Controlled access sources tended to provide entire datasets (79.5%), whereas open access more often supplied only main analysis variables (61.3%). This is reflected in the median number of explored variables by type of access: 449 for controlled access vs. 92 for open access. Datasets were provided in multiple software formats (22.9%), followed by SAS (21.4%) and CSV (17.1%). The most common associated documentation[10, 32] were study protocol (78.6%) and data dictionary (75.7%), with 48.6% including additional documentation such as patient information sheets, consent forms and supplementary tables.

Characteristics of the studies associated to the included datasets.

Table 2 presents a summary of the observed characteristics of the studies associated with the included datasets. Clinical trial registrations were found for 92% of the studies. Most registrations occurred from 2006 to 2015 (54.3%), and the studies were primarily published between 2011 and 2020 (70.0%). Most studies were multicentred (77.1%) and many were multinational (32.9%), involved adult participants (80%) and utilised a parallel design (82.9%). Studies were primarily from clinical trial phase III (38.6%) and 'not applicable' (35.7%). The median number of participants was 355. We included three studies related to rare diseases (4.3%) as defined by Orpha.net[33].

Re-identification risk scores results

Table 3 shows that the most common indirect identifiers in the datasets were age (84.3%) and sex (80.0%), followed by weight (47.1%) and height (44.3%). Nine (12.9%) datasets were automatically risk scored to 0 or 1. Datasets had a mean of 4 identifiers. Mean risk scores ranged from 0.47 to 0.91. Rb was usually higher than Ra (at all thresholds) and Rc. Moreover, the more indirect identifiers, the higher the risks scores (Appendix 6).

To further explore these results, pre-specified plots in Appendix 6 were used. Ras, Rb, and Rc did not seem to be correlated with the number of participants. While the risk scores provided distinct aspects of the dataset's granularity, a correlation between Ras and Rc was noticed; as the threshold increased, the correlation became stronger (Appendix 6). Table 4 shows the re-identifications risk scores from Table 3 categorised.

We did not encounter any reportable critical issues with the analysed datasets; hence, there was no need for us to communicate with any of the data owners or holders.

Re-identification risk scores in action

To understand the behaviour of the risk scores, we conducted two exploratory comparisons on three of the acquired 70 datasets.

First: We used the TOPPIC[34] trial anonymised dataset and calculated its risk score according to our protocol. We obtained 240 unique strata, matching its number of participants, and all risk scores were 1. Next, we categorised its continuous indirect identifiers (age, height and weight) in bands of 10 units. This yielded 123 unique strata, with risk scores ranging from 0.27 to 1. We then grouped them into bands of 20 units, resulting in 52 unique strata and risk scores ranging from 0.07 to 1. Then, we collapsed bands with counts less than 5

with their most adjacent band, producing 44 unique levels, with risk scores ranging from 0.05 to 1. From this last categorisation, we also removed age, then reinstated it, and subsequently excluded weight. In both cases, the number of strata were 17 and 15 respectively, with risk scores from 0.01 to 1 and 0.004 to 1 (Appendix 6).

Second: We compared risk calculations for two datasets (Appendix 6). Although both datasets have three indirect identifiers, all risk scores are higher for the IST trial[35] dataset (19435 participants and 2570 unique strata) compared to the RESTART trial[36] dataset risk scores (537 participants – 8 unique levels), as the former is more granular than the latter.

Discussion

The experience of dataset request and extraction

Securing 76 out of 86 (88.4%) requested datasets from 15 data sources, indicates a widespread willingness to share data, which is reassuring. Although our affiliation with a reputable academic institution likely improved our chances of securing them. Remarkably, we were not charged any fees by the data holders, which is encouraging given the evolving and increasingly strict regulatory environment for processing personal data[37, 38]. This situation may change in the future as data holders might start charging for the extra work involved[30, 39].

The time to obtain most open access datasets was short, 30 datasets received in 0 days, while one outlier taking 241 days, which the owner agreed to provide almost immediately, but it took 241 days to locate and send it. In contrast, procuring controlled access dataset, which required DSAs, was a lengthy and arduous process. Multiple forms needed to be completed before reaching the DSA stage, which often involved extensive negotiations between legal departments. This process should be reviewed, as it should not take nearly nine months to a year to obtain the datasets.

We observed, as other researchers[40] before us, that controlled access is not a universal concept, but involves a variety of processes. At the simplest end, we only had to fill out a request form, submit our CVs, and outline our research question (9/39 datasets, 23.1%). The next level of complexity involved signing DSAs alongside the initial steps (16/39 data requests, 41.0%). The most complicated process entailed the additional step of using the shared data in a TRE (14/39 data requests, 35.9%).

The characteristics of the data packs

Certain characteristics slowed down our analysis: unclear or unavailable data dictionaries, variables repeated multiple times requiring consistency checks, and outdated software formats. Conversely, some datasets following the CDISC Study Data Tabulation Model[41] or Analysis Data Model[42] expedited the analysis. Datasets should be accompanied by clear data dictionaries, avoid duplicated variables, and be stored in basic formats like CSV or TXT to avoid compatibility issues, or adhere to recognised standards.

The Interpretation of re-identification risk scores

We found that 10% (7/70) of the datasets were labelled as anonymised yet they contained direct identifiers (personal details), such as date of birth and participants' initials. These datasets were included in the analysis to highlight this issue, with all their risk scores automatically set to 1, representing the worst-case scenario. Researchers must carefully cross-check their anonymised datasets with the list provided by Hrynaszkiewicz et al.[27] before releasing them. Certain dataset characteristics increased the risk scores, for instance we encountered exact ages (e.g., measured in days) and dates of randomisation, which could be used to reverse-engineer dates of birth. These details should be removed as they are not necessary for analysis. Some datasets included the 'Date of Death' for participants; under GDPR, personal data protection is not applicable once an individual is deceased. However, Hrynaszkiewicz et al.[27] recommend removing all dates unique to a participant, as this could potentially affect the living relatives of the participants. Using the number of days from randomisation to death instead of exact dates of death retains analytical value while protecting privacy.

Notably, 2.9% (2/70) of the datasets had no identifiers or only one indirect identifier, resulting in risk scores automatically set to zero. Datasets with no identifiers could be freely shared without privacy implications and are a viable option for researchers. However, even if one (or no) indirect identifier remains, a holistic check must be made. For example, one obtained dataset recorded only age, but its publication indicated that all participants were female and located in a specific region of the UK.

Each re-identification risk score assesses different aspects of dataset granularity, so we cannot recommend any over the others. Furthermore, we cannot comment on their absolute magnitude as there are no standards or examples for comparison in clinical trials. However, smaller scores generally indicate better privacy protection[14]. Regarding Ra, statistical disclosure control[43] suggests suppressing table cells with counts less than five (i.e. a threshold of 0.2). Releasing datasets at this threshold could be an option, but it might reduce data usability. Researchers need to set a threshold that balances their data utility and privacy

risk. More importantly, it is crucial to acknowledge that the threshold cannot ever be zero if indirect identifiers are present. This understanding emphasises that anonymisation is a spectrum, not a binary state, and some risk must be endured[30]. Rb behaved like a discrete variable, often scoring 1 with few exceptions, regardless the number of identifiers. This outcome was expected since most included datasets had at least one stratum represented by only one participant. In the TOPPIC trial dataset comparison, Rb stubbornly remained at 1 despite attempts to reduce granularity. Lowering Rb[30] requires more sophisticated approaches like k-anonymity[44], where the smallest membership per stratum should be k. Clinical trials researchers might need support in learning new techniques to address this. Finally, Rc showed that datasets not only had a few problematic strata (i.e., with 1 or 2 participants) but also many strata with low participant membership.

Using re-identification risk scores

We recommend the process outlined in Figure 1 (adapted from El-Emam[45]) to effectively use the risk scores. First, remove all direct identifiers, followed by verifying the presence of indirect identifiers. Next, question the necessity of these indirect identifiers to maintain utility[46]. Note, there is no universally agreed-upon interpretation of data utility, it is context-specific. Researchers must define what utility means for their study. The ability to reproduce the primary outcome analysis with the anonymised dataset can serve as a measure of utility, as shown by others [40] [46].The researcher's definition of utility[47] is pivotal in shaping the anonymised datasets. For example, the TOPPIC[34] trial dataset was designed to allow investigation of new research questions (423 variables, 240 participants, 240 unique levels with four indirect identifiers, Appendix 6) while the RESTART[36] trial dataset was designed from primary analysis replication (66 variables, 537 participants, eight unique levels with three indirect identifiers, Appendix 6). Once utility is defined, calculate the risk scores, if they are considered high, they could be lowered by manipulating the data through perturbation, recalculation, recoding and suppression[11] and by using privacy models[44] [48] [49] [50]. Finally, use Table 4 to compare the newly anonymised dataset's re-identification risk scores with those from the datasets included in this research and decide the type of access to be used for release.

Strengths and limitations

This study covers an emerging area in clinical trials methods research, where even the definition of anonymisation lacks consensus. Therefore, our selection of datasets defined as anonymised or de-identified, may have introduced heterogeneity among the characteristics of the requested datasets.

The available datasets currently underrepresent Africa and Oceania, which might limit the applicability of our findings to those continents. To enhance the representation of open access datasets, we modified the inclusion criteria for datasets from BMJ and PlosOne. This may have skewed our findings and affected the analysis publication dates, as our search was limited to datasets from January 2015 onward for those data sources.

We did not find, or actively seek, trials studying risky or stigmatising behaviour, which would require special treatment beyond the scope of this research[51-53]. Furthermore, the journalist re-identification risk scores might have been overestimated, as all the matching theoretical datasets were 15 times larger than the original dataset. This scale may not have been sufficient in some case, especially for very small datasets (fewer than 40 participants). Moreover, the probability of the existence of a real-life matching dataset was not evaluated, as this requires specialised expertise.

At least 19 open access datasets were available when we wrote the protocol for this research, making the project feasible. However, we could not have predicted that controlled access repositories would be as receptive to our requests as they were. In hindsight, with greater ambition, we could have requested more datasets, thereby increasing our sample size, but given limited resources, it would not have increased significantly.

The re-identification risk scores alone do not determine the vulnerability of datasets to a re-identification attack in the real world[54, 55] [56] [57], it is influenced by factors such as an attacker's motivations, resources, and potential gains, which were outside the scope of this research.

There is potential to quickly start the implementation of risk scores calculation for clinical trial dataset using the freely available R package sdcMicro[58-60]. The main strength of this research is that it brought together a variety of datasets and assessed them using a simple methodology under the same conditions.

Conclusion

This study confirms a strong inclination to share clinical trial datasets, which are rich in personal details. Although re-identification risk scores may appear oversensitive for clinical trial datasets, which are typically smaller than datasets from electronic health records, their high magnitudes do not necessarily translate into real-life re-identification threats for participants. Instead, these scores provide valuable summaries for

343 understanding the granularity of clinical trial datasets, aiding decision-making for dataset release for
344 secondary purposes.
345 We have demonstrated that calculating re-identification risk scores is simple and feasible. The number and
346 type of identifiers (continuous or discrete) are crucial in controlling risk scores. More identifiers increase
347 granularity, with continuous identifiers adding more granularity than discrete ones. While risk scores alone
348 cannot determine if data is sufficiently anonymised or protected, they can assist in calibrating the
349 anonymisation process of clinical trial datasets.
350 Researchers can use our findings to assess how their anonymised datasets compare with ours. Clinical trial
351 researchers should consider employing the process outlined in this research to inform their anonymisation
352 procedures before releasing data. The proposed method is a cost-effective and pragmatic solution. While
353 simple, it is highly informative and serves as a useful stopgap, especially when resources are limited.
354

355 **Declarations**

356 **Ethics and dissemination**

357 Prior to commencement, the research was subject to the University of Edinburgh's Usher Institute
358 ethics/data protection oversight process. The ethics/data protection triage and overview self-audit of
359 ethics/data protection issues, completed on December 3, 2020 (by AR and SL Appendix 7), confirmed
360 that the proposed research (being fully anonymous secondary data analysis) posed no reasonably
361 foreseeable ethics/data protection risks. This indicated that there was no requirement for proceeding to
362 full formal ethics/data protection review by the Usher Research Ethics Group.

363 **Availability of data and materials**

364 Anonymised raw re-identification risk data and its analysis code may be requested from the corresponding
365 author for further reasonable research.

366 **Conflicts of Interests**

367 The authors declare no competing interests.

368 **Funding**

369 AR has a scholarship from the University of Edinburgh to undertake a PhD with the support from the Asthma
370 UK Centre for Applied Research (AUKCAR grant no. AUK-AC-2012-01).
371 SCL, LJW, PS and CJW are supported in this work by their employment at the Edinburgh Clinical Trials Unit.

372 TJ is supported by Asthma UK as part of the Asthma UK Centre for Applied Research (grant nos. AUK-AC-
373 2012-01 and AUK-AC-2018-01).
374 SE is supported in this work by her employment at the Pragmatic Clinical Trials Unit.
375 All of the authors contributed to protocol and manuscript development. Neither sponsor (AUKCAR) nor
376 funder (University of Edinburgh) contributed to protocol or manuscript development.

379 **Author contributions**

380 AR, SCL and CJW conceived the idea for this work supported by SE. AR wrote the first draft and all authors
381 contributed to this article.

382 **Acknowledgements**

383 Thanks to all data holder/owners for providing access to the data and to all clinical trial participants who
384 permitted the use of their data for secondary research. Requested acknowledgments statements were:

id	Source (included studies)	Requested acknowledgments statements
1	https://datacompass.lshtm.ac.uk [61] (NCT02104232[62], NCT02111915[63], ISRCTN36436933[64])	No statement required.
2	https://ctu-app.lshtm.ac.uk/freebird [65] (ISRCTN7445979[66], NCT00375258[67], NCT00872469[68] NCT00872469[69], NCT03777488[70])	No statement required
3	https://datashare.is.ed.ac.uk [71] (ISRCTN45178534[72] , ISRCTN25765518[73], IST (Registration not required)[35], ISRCTN71907627[36] ISRCTN89489788[34])	we gratefully acknowledge the IST-3 Collaborative Group, the trial joint sponsors (The University of Edinburgh and the Lothian Health Board), and the chief funding agencies of the study: UK Medical Research Council, Health Foundation UK, Stroke Association UK, Research Council of Norway, Arbetsmarknadens Partners Forsakringsbolag (AFA) Insurances Sweden, Swedish Heart Lung Fund, The Foundation of Marianne and Marcus Wallenberg, Polish Ministry of Science and Education, the Australian Heart Foundation, Australian National Health and Medical Research Council (NHMRC), Swiss National Research Foundation, Swiss Heart Foundation, Assessorato alla Sanita, Regione dell'Umbria, Italy, and Danube University
4	https://www.clinicalstudydatarequest.com [74] (Registration not required[75] , NCT01822899[76] NCT01842607[77], NCT01405053[78], NCT00948766[79])	This publication is based on research using data from the Sponsor companies GlaxoSmithKline Research & Development Ltd, Eisai Limited and Novartis Pharma AG that have been made available to us through secured access. CSDR team or Sponsors have not contributed to or approved, and are not in any way responsible for, the contents of this publication. We thank both Sponsors and CSDR for providing us data and access.
5	http://datadryad.org [80] (ACTRN12616000888460[81], HKCTR-1848[82] Registration no required[83], NCT04523831[84])	No statement required
6	http://yoda.yale.edu [22] (NCT01715285[85], NCT00903331[86] NCT01004432[87], NCT00211133[88])	This study, carried out under YODA Project 2022-4951, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C.. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C..

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

id	Source (included studies)	Requested acknowledgments statements
7	https://www.projectdatasphere.org [21] (NCT00058474[89], NCT00033293[90], NCT00310180[91], NCT00693992[92], NCT00312208[93],NCT00143455[94] NCT00113763[95],NCT00617669[96], NCT00676650[97])	This publication is based on information obtained from https://data.projectdatasphere.org, which is maintained by Project Data Sphere, and includes information that has been made available by the National Cancer Institute and is also available through the National Clinical Trials Network (NCTN)/NCI Community Oncology Research Program (NCORP) Data Archive. The information was collected from the following clinical trials: • A Clinical Trial Comparing Preoperative Radiation Therapy And Capecitabine With or Without Oxaliplatin With Preoperative Radiation Therapy And Continuous Intravenous Infusion Of 5-Fluorouracil With or Without Oxaliplatin In The Treatment Of Patients With Operable Carcinoma Of The Rectum. NCT00058474. • A Pilot Study Randomized Trial of Intravenous Gammaglobulin Therapy for Patients With Neuroblastoma Associated Opsoclonus-Myoclonus-Ataxia Syndrome Treated With Chemotherapy and Prednisone. NCT00033293. • Program for the Assessment of Clinical Cancer Tests (PACCT-1): Trial Assigning Individualized Options for Treatment:The TAILORx Trial. NCT00310180. • Randomized, Phase III, Double-Blind Placebo-Controlled Trial of Sunitinib (NSC #736511) as Maintenance Therapy in Non-progressing Patients Following an Initial Four Cycles of Platinum-Based Combination Chemotherapy in Advanced, Stage IIIB / IV Non-small Cell Lung Cancer. NCT00693992. • A Multicenter Phase III Randomized Trial Comparing Docetaxel in Combination With Doxorubicin and Cyclophosphamide Versus Doxorubicin and Cyclophosphamide Followed by Docetaxel as Adjuvant Treatment of Operable Breast Cancer HER2neu Negative Patients With Positive Axillary Lymph Nodes. NCT00312208. • Open Label, Randomised Multicentre Phase III Study Of Irinotecan Hydrochloride (Campto (Registered)) And Cisplatin Versus Etoposide And Cisplatin In Chemotherapy Naive Patients With Extensive Disease - Small Cell Lung Cancer. NCT00143455. • An Open-label, Randomized, Phase 3 Clinical Trial of ABX-EGF Plus Best Supportive Care Versus Best Supportive Care in Subjects With Metastatic Colorectal Cancer. NCT00113763. • A Phase III, Randomised, Double-blind, Placebo-controlled Study to Assess the Efficacy and Safety of 10 mg ZD4054 (Zibotentan) in Combination With Docetaxel in Comparison With Docetaxel in Patients With Metastatic Hormone-resistant Prostate Cancer. NCT00617669. • A Multicenter, Randomized, Double-Blind, Phase 3 Study Of Sunitinib Plus Prednisone Versus Prednisone In Patients With Progressive Metastatic Castration-Resistant Prostate Cancer After Failure Of A Docetaxel-Based Chemotherapy Regimen. NCT00676650. All analyses and conclusions in this publication are the sole responsibility of the authors and do not necessarily reflect the opinions of the owners of the information, the clinical trial investigators, the NCTN, the NCORP, the National Cancer Institute, or Project Data Sphere. Neither the owners of the information, the clinical trial investigators, the NCTN, the NCORP, the National Cancer Institute, nor Project Data Sphere have contributed to, approved or are in any way responsible for the contents of this publication
8	https://biolincc.nhlbi.nih.gov/studies [98] (NCT00650091[99], NCT00000589[100], NCT00075829[101], NCT01982968[102], NCT01134783[103], NCT00004562[104])	This Manuscript was prepared using BMTCTN0102, CHOICES, PANTHER, OAT, TRAP, WRAP_IPF Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the BMTCTN0102, CHOICES, PANTHER, OAT, TRAP, WRAP_IPF or the NHLBI
9	https://nda.nih.gov/get/access-data.html [105] (NCT00012558[106], Registration not found[107], NCT01927276[108], NCT01944046[109],NCT00005013[110])	Data and/or research tools used in the preparation of this manuscript were obtained from the National Institute of Mental Health (NIMH) Data Archive (NDA). NDA is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in mental health.

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

id	Source (included studies)	Requested acknowledgments statements
		Dataset identifier(s): 3058, 2147, 2622, 2724, 2009, 2157. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or of the Submitters submitting original data to NDA.
10	https://vivli.org/ [111] (NCT01198756[112], NCT01573767[113], NCT01313676[114], NCT00400855[115], NCT01498822[116])	This publication is based on research using data from data contributors GSK and UCB that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication.
11	https://beta.ukdataservice.ac.uk/datacatalogue/studies (reshare.ukdataservice.ac.uk) (https://www.ukdataservice.ac.uk/deposit-data) [117] (ISRCTN11288961[118], ISRCTN90749868[119] NCT01801410[120], Registration not required[121], ISRCTN24081411[122])	To Nottingham Clinical Trials Unit for facilitating sharing of data from the PRIDE study via ukdataservice.ac.uk
12	https://med.data.edu.au/find-data/ [123]	Repository no longer available
13	https://dcri.org/our-approach/data-sharing/soar-data [124] SOAR data: Available datasets: Duke cardiac catheterization datasets.	We acknowledge Michael Cohen-Wolkowiez, who is the principal investigator who conducted the original study from which the data were generated. Furthermore, we acknowledge the Eunice Kennedy Shriver National Institute of Child Health and Human Development Data and Specimen Hub for providing the Pharmacokinetics of Clindamycin and Trimethoprim-sulfamethoxazole in Infants and Children (PBPK) data that were used for this research.
14	https://journals.plos.org/plosone/search [125] (NCT02700490[126], TCTR20201005002[127], NCT02185196[128], ACTRN12616000538448[129], ISRCTN 71217488[130] NCT02747524[131])	No statement required
15	https://www.bmj.com/search/advanced [132] (NCT02068885[133], ,NCT01953549[134], ISRCTN11980540[135])	No statement required
16	https://dataverse.harvard.edu/ [136] (CTRI/2016/09/007240[137], PACTR201901905832601[138] NCT02148952[139], SLCTR/2019/015[140], ANZCTR126160013674 37[141])	No statement required
17	https://arlg.org/studies-in-progress/ [142]	Not applicable as RCT data was not located in this repository
18	https://repository.niddk.nih.gov/studies/ [143]	Not applicable as RCT data was not located in this repository

385

386

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

387 ***Tables***

Table 1 - **Datasets' Characteristics**

Parameter	Category/ Description	Controlled N=39	Open N=31	Overall N=70
Access n(%)	Controlled-Vetoing	9 (23.1)	--	9 (12.8)
	Controlled-Vetoing + DSA	16 (41.0)	--	16 (22.9)
	Controlled-Vetoing + DSA + TRE	14 (35.9)	--	14 (20.0)
	Open - No restrictions	--	31 (100)	31 (44.3)
Days to obtain	mean (sd)	270 (125)	8 (43)	154 (163)
	median (IQR)	238 (231-343)	0 (0-0)	138 (0-238)
Type dataset (n(%))	Main analysis	7 (17.9)	19 (61.3)	26 (37.1)
	Complete trial	31 (79.5)	7 (22.6)	38 (54.3)
	Not enough information	1 (2.6)	5 (16.1)	6 (8.6)
No of variables explored (n(%))	mean (sd)	1610 (4076)	160 (211)	968 (3113)
	median (IQR)	449 (217-918)	92 (39-217)	237 (56-572)
File format (n(%))	CSV	6 (15.4)	6 (19.4)	12 (17.1)
	SAS	13 (33.3)	2 (6.5)	15 (21.4)
	SPSS	--	3 (9.7)	3 (4.3)
	STATA	2 (5.1)	3 (9.7)	5 (7.1)
	TXT	5 (12.8)	--	5 (7.1)
	XLSX	--	10 (32.3)	10 (14.3)
	XPT	4 (10.3)	--	4 (5.7)
	Multiple[1]	9 (23.1)	7 (22.6)	16 (22.9)
Data dictionary n(%)	Available	35 (89.7)	18 (58.1)	53 (75.7)
Case report form n(%)	Available	23 (59.0)	7 (22.6)	30 (42.9)
Protocol n(%)	Available	33 (84.6)	22 (71.0)	55 (78.6)
Statistical analysis plan n(%)	Available	16 (41.0)	7 (22.6)	23 (32.9)
Clinical summary report n(%)	Available	11 (28.2)	--	11 (15.7)
Anonymisation details[2] n(%)	Available	13 (33.3)	--	13 (18.6)
Other documents[3] n(%)	Available	17 (43.6)	20 (64.5)	34 (48.6)

Notes:	1 Multiple refers to two or more of the above formats were available for a single dataset 2 Documentation detailing how the anonymisation was executed 3 Other documents refer to patient information sheets, ethical approval letters, standard operating procedures, editorials, summary tables, lay summaries, supplements, consent forms, evaluation instrument manuals
Where	DSA: Data sharing agreement sd: Standard deviation IQR: Inter quartile range TRE: Trusted research environment N: Number of datasets n: Number of observations

388

389

Table 2 - **Associated Studies Characteristics**

Parameter	Category/ Description	Controlled N=39	Open N=31	Overall N=70
Year of registration n(%)	Pre 2000	3 (7.7)	1 (3.2)	4 (5.7)
	2001-2005	7 (17.9)	3 (9.7)	10 (14.3)
	2006-2010	13 (33.3)	2 (6.5)	15 (21.4)
	2011-2015	13 (33.3)	10 (32.3)	23 (32.9)
	2016-2020	1 (2.6)	12 (38.7)	13 (18.6)
	Not required[1]	1 (2.6)	3 (9.7)	4 (5.7)
	Not found/available	1 (2.6)	--	1 (1.4)
Year published n(%)	Pre 2000	2 (5.1)	1 (3.2)	3 (4.3)
	2001-2005	2 (5.1)	1 (3.2)	3 (4.3)
	2006-2010	--	1 (3.2)	4 (5.7)
	2011-2015	12 (30.8)	4 (12.9)	16 (22.9)
	2016-2020	16 (41.0)	17 (54.8)	33 (47.1)
	2021 and after	3 (7.7)	7 (22.6)	10 (14.3)
	Not published	3 (xx)	--	1 (1.4)
Sites n(%)	Multicentre	37 (94.9)	17 (54.8)	54 (77.1)
	Single	1 (2.6)	14 (45.2)	15 (21.4)
	Not available	1 (2.6)	--	1 (1.4)
Location n(%)	Africa	--	3 (9.7)	3 (4.3)
	Asia	3 (7.7)	10 (32.3)	13 (18.6)
	Europe and UK	3 (7.7)	7 (22.6)	8 (14.3)
	Multinational	18 (46.2)	5 (16.1)	23 (32.9)
	Oceania	1 (2.6)	3 (9.7)	4 (5.7)
	USA and Americas	14 (35.9)	3 (9.7)	19 (24.3)
Population n(%)	Adults	29 (74.4)	27 (87.1)	56 (80.0)
	Adults and children	4 (10.3)	1 (3.2)	5 (7.1)
	Children	6 (15.4)	3 (9.7)	9 (12.9)
Design n(%)	Cluster	1 (2.6)	4 (12.9)	5 (7.1)
	Crossover	3 (7.7)	1 (3.2)	4 (5.7)
	Factorial	1 (2.6)	1 (3.2)	2 (2.9)
	Parallel	33 (84.6)	25 (80.6)	58 (82.9)
	SMART	1 (2.6)	--	1 (1.4)
Phase n(%)	I	--	2 (6.5)	2 (2.9)
	II	4 (10.3)	1 (3.2)	5 (7.1)
	III	21 (53.8)	6 (19.4)	27 (38.6)
	IV	2 (5.1)	1 (3.2)	3 (4.3)
	Not applicable	10 (25.6)	15 (48.4)	25 (35.7)
	Not available	2 (5.1)	6 (19.4)	8 (11.4)
Rare disease[2] n(%)	Yes	3 (7.7)	--	3 (4.3)
Number of participants	mean(sd)	1584 (4064)	7952 (29574)	4404 (19988)
	median (IQR)	470 (240-939)	240 (64-903)	355 (154-921)

390

Notes:	1 Not required refers to studies too old (pre implementation of registration) or not in the remit for registration 2 As classified by orpha.net
Where	N: Number of datasets n: Number of observations

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Table 3 – **Identifiers and Re-identification risks scores for included datasets**

Parameter	Category/ Description	Controlled N=39	Open N=31	Overall N=70
Indirect Identifiers (n(%))[1]	Age	29 (74.4)	22 (70.1)	51 (72.9)
	Age category	7 (17.9)	1 (3.2)	8 (11.4)
	Country	14 (35.9)	4 (12.9)	18 (25.7)
	Education	5 (12.8)	7 (22.6)	12 (17.1)
	Ethnicity	20 (51.3)	4 (12.9)	25 (35.7)
	Race	27 (69.2)	2 (6.5)	28 (40.8)
	Sex	37 (94.9)	21 (67.7)	56 (80.0)
	Marital Status	4 (10.3)	1 (3.2)	5 (7.1)
	Occupation	4 (10.3)	1 (3.2)	5 (5.7)
	Height	25 (64.1)	6 (19.4)	31 (44.3)
	Weight	26 (66.7)	7 (22.6)	33 (47.1)
	Date of randomisation	--	3 (9.7)	3 (4.3)
	Others[2]	6	13	19
Number of identifiers (n(%))	0[3]	--	2 (6.5)	2 (2.9)
	1[4]	1 (2.6)	6 (19.4)	7 (10.0)
	2	2 (5.1)	5 (16.1)	7 (10.0)
	3	2 (5.1))	8 (25.8)	10 (14.3)
	4	10 (25.6)	2 (6.5)	12 (17.1))
	5	7 (17.9)	4 (12.9)	11 (15.7)
	6	6 (15.4))	3 (9.7)	9 (12.9)
	7	10 (25.6)	1 (3.2)	11 (15.7)
	8	1 (2.6)	--	1 (1.4)
Number of identifiers	mean(sd)	5.1 (1.67)	3.0 (1.9)	4.2 (2.06)
Prosecutor Ra (mean (sd))	Threshold at 0.01	0.92 (0.24)	0.86 (0.32)	0.89 (0.28)
	Threshold at 0.05	0.87 (0.31)	0.82 (0.37)	0.85 (0.34)
	Threshold at 0.1	0.84 (0.33)	0.80 (0.38)	0.82 (0.36)
	Threshold at 0.2	0.81 (0.34)	0.76 (0.40)	0.79 (0.37)
	Threshold at 0.3	0.79 (0.35)	0.75 (0.40)	0.77 (0.37)
	Threshold at 0.4	0.76 (0.35)	0.72 (0.41)	0.75 (0.37)
	Threshold at 0.5	0.70 (0.37)	0.65 (0.41)	0.68 (0.39)
Prosecutor Rb (mean (sd))	--	0.93 (0.24)	0.88 (0.33)	0.91 (0.28)
Prosecutor Rc mean (sd))	--	0.76 (0.34)	0.71 (0.39)	0.74 (0.36)
Journalist Ra (mean (sd))	Threshold at 0.01	0.71 (0.31)	0.81 (0.26)	0.75 (0.29)
	Threshold at 0.05	0.69 (0.30)	0.80 (0.26)	0.74 (0.28)
	Threshold at 0.1	0.60 (0.31)	0.70 (0.31))	0.65 (0.31)
	Threshold at 0.2	0.53 (0.35)	0.61 (0.34)	0.57 (0.34)
	Threshold at 0.3	0.50 (0.35)	0.57 (0.36)	0.53 (0.35)
	Threshold at 0.4	0.48 (0.36)	0.52 (0.37)	0.50 (0.36)
	Threshold at 0.5	0.46 (0.36)	0.48 (0.38)	0.47 (0.37)
Journalist Rb (mean (sd))	--	0.83 (0.35)	0.80 (0.38)	0.82 (0.36)
Journalist Rc (mean (sd))	--	0.73 (0.41)	0.57 (0.43)	0.66 (0.43)

Notes:
1 These are not mutually exclusive categories; a dataset could have more than one of these indirect identifiers. Therefore, percentages would not add to 100%
2 Other identifiers refer to, BMI, deprivation index, number of children in the family, living siblings, number of pregnancies, education of parents, religion, minority, student status, city, location name, income, socio economic status, all with a count of 1, but a dataset could have more than one of these indirect identifiers, therefore percentage is not applicable
3 One dataset has no identifiers and 1 dataset has 1 identifier 'sex', both datasets' risks coded automatically to zero
4 All with a direct identifier, all datasets' risks coded automatically to one

Where:
Risk a (Ra): the proportions of participants in strata above a predetermined risk threshold, Risk b (Rb): the stratum with the smallest membership in the anonymised dataset, and Risk c (Rc): the average risk score across the whole strata of the anonymised dataset, using all indirect identifiers, calculated with the formulas in chapter 16 from "Guide to the de-identification of personal health information" by Khaled El Emam (2013)
N: Number of datasets
n: Number of observations

391

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Table 4 - **Re-identification risks scores categorised for included datasets**

		Prosecutor			Journalist		
	Risk interval	Controlled N=39	Open N=31	Overall N=70	Controlled N=39	Open N=31	Overall N=70
Ra at 0.01	0	1 (2.6)	2 (6.5)	3 (4.3)	--	2 (6.5)	2 (2.9)
(n(%))	0> and <=0.25	1 (2.6)	1 (3.2)	2 (2.9)	5 (12.8)	--	5 (7.1)
	0.25> and <=0.5	2 (5.1)	2 (6.5)	4 (5.7)	3 (7.7))	1 (3.2)	4 (5.7)
	0.5> and <=0.75	--	--	--	4 (10.3)	5 (16.1)	9 (12.9)
	0.75> and <1	--	1 (3.2)	1 (1.4)	23 (59.0)	14 (45.2)	37 (52.9)
	1	35 (89.7)	25 (80.6)	60 (85.7)	4 (10.3)	9 (29.0)	13 (18.6)
Ra at 0.05	0	1 (2.6)	3 (9.7)	4 (5.7)	--	2 (6.5)	2 (2.9)
(n(%))	0> and <=0.25	4 (10.3)	2 (6.5)	6 (8.6)	5 (12.8)	--	5 (7.1)
	0.25> and <=0.5	--	1 (3.2)	1 (1.4)	3 (7.7)	1 (3.2)	4 (5.7)
	0.5> and <=0.75	1 (2.6)	--	1 (1.4)	5 (12.8)	5 (16.1)	10 (14.3)
	0.75> and <1	4 (10.3)	1 (3.2)	5 (7.1)	23 (59.0)	15 (48.4)	38 (54.3)
	1	29 (74.4)	24 (77.4)	53 (75.7)	3 (7.7)	8 (25.8)	11 (15.7)
Ra at 0.1	0	2 (5.1)	3 (9.7)	5 (7.1)	1 (2.6)	3 (9.7)	4 (5.7)
(n(%))	0> and <=0.25	3 (7.7)	2 (6.5)	2 (2.9)	6 (15.4)	1 (3.2)	7 (10.0)
	0.25> and <=0.5	1 (2.6)	2 (6.5)	3 (4.3)	6 (15.4)	1 (3.2)	7 (10.0)
	0.5> and <=0.75	1 (2.6)	--	1 (1.4)	6 (15.4)	8 (25.8)	14 (20.0)
	0.75> and <1	5 (12.8)	1 (3.2)	6 (8.6)	18 (46.2)	12 (38.7)	30 (42.9)
	1	27 (69.2)	23 (74.2)	50 (71.4)	2 (5.1)	6 (19.4)	8 (11.4)
Ra at 0.2	0	2 (5.1)	4 (12.9)	6 (8.6)	4 (10.3)	3 (9.7)	7 (10.0)
(n(%))	0> and <=0.25	3 (7.7)	3 (9.7)	6 (8.6)	7 (17.9)	2 (6.5)	9 (12.9)
	0.25> and <=0.5	2 (5.1)	--	2 (2.9)	4 (10.3)	4 (12.9)	8 (11.4)
	0.5> and <=0.75	--	2 (6.5)	2 (2.9)	8 (20.5)	10 (32.3)	18 (25.7)
	0.75> and <1	10 (25.6)	6 (19.4)	16 (22.9)	14 (35.9)	6 (19.4)	20 (28.6)
	1	22 (56.4)	16 (51.6)	38 (54.3)	2 (5.1)	6 (19.4)	8 (11.4)
Ra at 0.3	0	3 (7.7)	4 (12.9)	7 (10.0)	4 (10.3)	4 (12.9)	8 (11.4)
(n(%))	0> and <=0.25	3 (7.7)	3 (9.7)	6 (8.6))	8 (20.5)	4 (12.9)	12 (17.1)
	0.25> and <=0.5	1 (2.6)	--	1 (1.4)	6 (15.4)	2 (6.5)	8 (11.4)
	0.5> and <=0.75	2 (5.1)	2 (6.5)	4 (5.7)	5 (12.8)	10 (32.3)	15 (21.4)
	0.75> and <1	10 (25.6)	7 (22.6)	17 (24.3)	14 (35.9)	5 (16.1)	19 (27.4)
	1	20 (51.3)	15 (48.4)	35 (50.0)	2 (5.1)	6 (19.4)	8 (11.4)
Ra at 0.4	0	3 (7.7)	4 (12.9)	7 (10.0)	4 (10.3)	5 (16.1)	9 (12.9)
(n(%))	0> and <=0.25	3 (7.7)	3 (9.7)	6 (8.6)	10 (25.6)	4 (12.9)	14 (20.0)
	0.25> and <=0.5	1 (2.6)	2 (6.5)	3 (4.3)	4 (10.3)	5 (16.1)	9 (12.9)
	0.5> and <=0.75	6 (15.4)	--	6 (8.6)	5 (12.8)	7 (22.6)	12 (17.1)
	0.75> and <1	6 (15.4)	8 (25.8)	14 (20.0)	14 (35.9)	4 (12.9)	18 (25.7)
	1	20 (51.3)	14 (45.2)	34 (48.6)	2 (5.1)	6 (19.4)	8 (11.4)
Ra at 0.5	0	3 (7.7)	4 (12.9)	7 (10.0)	5 (12.8)	5 (16.1)	10 (14.3)
(n(%))	0> and <=0.25	4 (10.3)	4 (12.9)	8 (11.4)	11 (28.2)	5 (16.1)	16 (22.9)
	0.25> and <=0.5	5 (12.8)	2 (6.5)	7 (10.0)	3 (7.7)	7 (22.6)	10 (14.3)
	0.5> and <=0.75	4 (10.3)	3 (9.7)	7 (10.0)	4 (10.3)	4 (12.9)	8 (11.4)
	0.75> and <1	11 (28.2)	5 (16.1)	16 (22.9)	14 (35.9)	4 (12.9)	18 (25.7)
	1	12 (30.8)	13 (41.9)	25 (35.7)	2 (5.1)	6 (19.4)	8 (11.4)
Rb	0	--	2 (6.5)	2 (2.9)	--	2 (6.5)	2 (2.9)
(n(%))	0> and <=0.25	3 (7.7)	2 (6.5)	5 (7.1)	6 (15.4)	4 (12.9)	10 (14.3)
	0.25> and <=0.5	--	--	--	2 (5.1)	1 (3.2)	3 (4.3)
	0.5> and <=0.75	--	--	--	--	--	--
	0.75> and <1	--	--	--	--	--	--
	1	36 (92.3)	27 (87.1)	63 (90.0)	31 (79.5)	24 (77.4)	55 (78.6)
Rc	0	--	2 (6.5)	2 (2.9)	--	2 (6.5)	2 (2.9)
(n(%))	0> and <=0.25	6 (15.4)	5 (16.1)	11 (15.7)	10 (25.6)	8 (25.8)	18 (25.7)
	0.25> and <=0.5	1 (2.6)	1 (3.2)	2 (2.9)	1 (2.6)	4 (12.9)	5 (7.1)
	0.5> and <=0.75	7 (17.9)	3 (9.7)	10 (14.3)	1 (2.6)	3 (9.7)	4 (5.7)
	0.75> and <1	13 (33.3)	7 (22.6)	20 (28.6)	2 (5.1)	1 (3.2)	3 (4.3)
	1	12 (30.8)	13 (41.9)	25 (35.7)	25 (64.1)	13 (41.9)	38 (54.3)

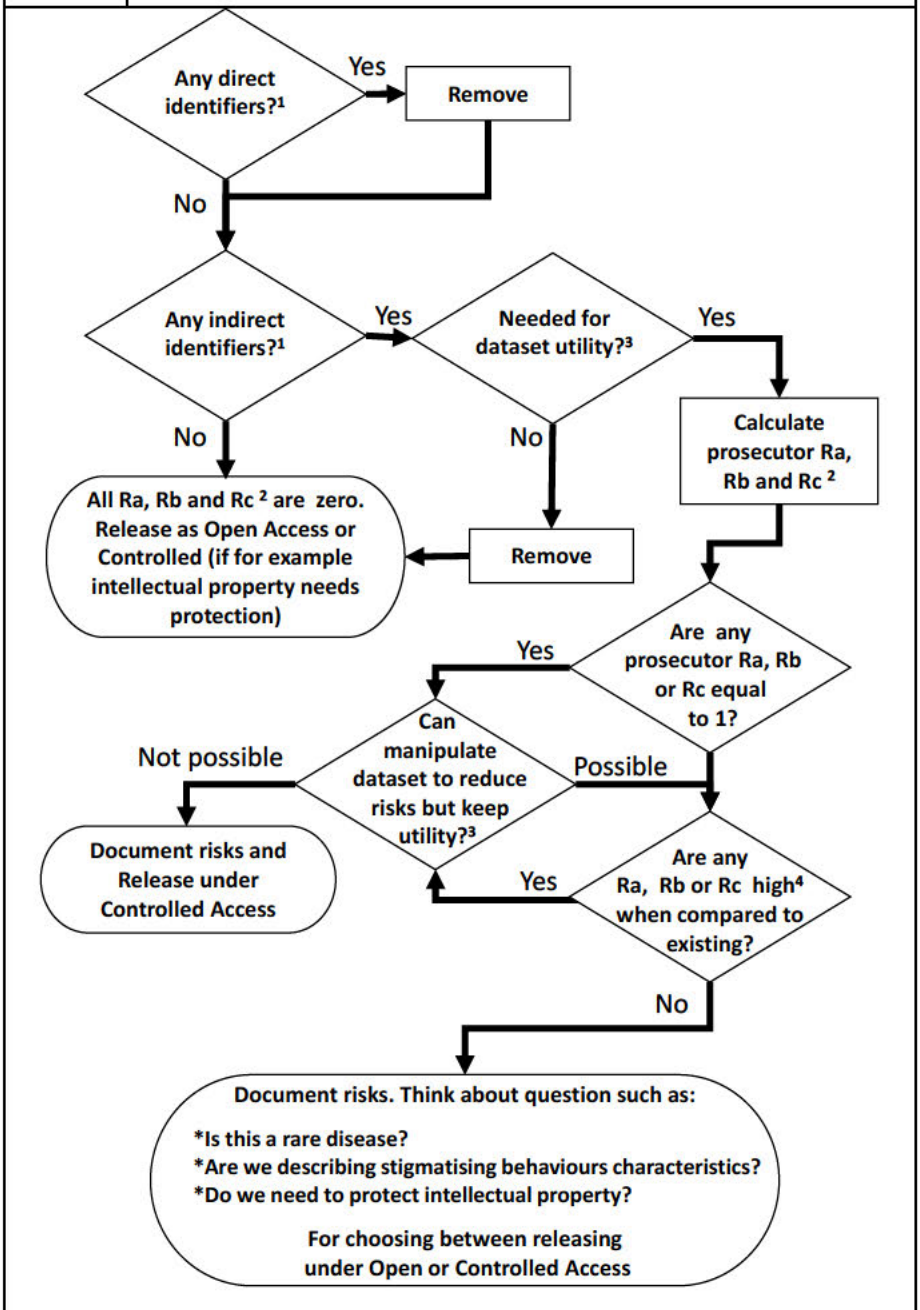
Where: Risk a (Ra): the proportions of participants in strata above a predetermined risk threshold, Risk b (Rb): the stratum with the smallest membership in the anonymised dataset, and Risk c (Rc): the average risk score across the whole strata of the anonymised dataset, using all indirect identifiers, calculated with the formulas in chapter 16 from "Guide to the de-identification of personal health information" by Khaled El Emam (2013)
N: Number of datasets
n: Number of observations

392

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

393 ***Figures***

Figure 1 Decision cycle for releasing clinical trials datasets using re-identification risks.



Where:

1 As described by Hrynaszkiewicz et al. (2015)

2 Risk a (Ra): the proportions of participants in strata above a predetermined risk threshold, Risk b (Rb): the stratum with the smallest membership in the anonymised dataset, and Risk c (Rc): the average risk score across the whole strata of the anonymised dataset, using all indirect identifiers, calculated with the formulas in chapter 16 from "Guide to the de-identification of personal health information" by Khaled El Emam (2013).

3 Utility is interpreted as the ability to reproduce the primary outcome analysis with the manipulated indirect identifiers

4 Using the parameters in figures 2 and 3 of this paper

394

Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendices

1	Appendix 1 Protocol	Currently a separate file
2	Appendix 2 Repositories	Currently a separate file
3	Appendix 3 Metadata collected	Currently a separate file
4	Appendix 4 Mock calculation of re-identification risks	Currently a separate file
5	Appendix 5 Datasets used	Currently a separate file
6	Appendix 6 Extra figures	Currently a separate file
7	Appendix 7 Ethical approval	Currently a separate file

Reference List

1. Dal-Ré R. Access to Anonymized Individual Participant Clinical Trials Data: A Radical Change of Mind by the Most Prestigious Medical Journals. *Archivos de Bronconeumologia* 2018; 54: 65-67. DOI: 10.1016/j.arbr.2017.12.007.
2. National Institutes of Health (NIH). Final NIH Policy for Data Management and Sharing. In: Sharing NSD, (ed.). USA2023.
3. Cancer Research UK. Data sharing guidelines, https://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy/data-sharing-guidelines (accessed 30 Oct 2020 2020).
4. Medical Research Council (MRC). MRC Policy on Open Research Data from Clinical Trials and Public Health Intervention Studies Updated 2019 ed. 2016.
5. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials. *BMJ* 2017; 357: j2372. DOI: 10.1136/bmj.j2372.
6. The EQUATOR Network. New ICMJE Recommendations published https://www.equator-network.org/2018/12/21/new-icmje-recommendations-published/ (2018).
7. Pisani E, Aaby P, Breugelmans JG, et al. Beyond open data: realising the health benefits of sharing data. *BMJ* 2016; 355: i5295. DOI: 10.1136/bmj.i5295.
8. Bertagnolli M, Sartor O, Chabner B, et al. Advantages of a truly open-access data-sharing model. *N Engl J Med* 2017; 12: 1178-1181. DOI: 10.1056/NEJMsb1702054.
9. Nadkarni P. Chapter 9 - Clinical data repositories: warehouses, registries, and the use of standards. *Clinical research computing: a practitioner's handbook*. Amsterdam: Elsevier, 2016, pp.173-185.
10. Tudur Smith C, Hopkins C, Sydes MR, et al. Good practice principles for sharing individual participant data from publicly funded clinical trials Version 1 ed.: Medical Research Council - Hubs for Trials Methodology Research, 2015.
11. Rodriguez A, Tuck C, Dozier MF, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. *Clinical Trials* 2022; 19: 452-463.
12. Walker N. All or Nothing: The False Promise of Anonymity. *bioRxiv* 2016: 084921.
13. Boris L. Re-identification of "anonymised data". *1 GEO L TECH REV 202* 2017: 202-213.
14. El Emam K. *Guide to the de-identification of personal health information*. CRC Press, 2013.
15. Dankar FK, El Emam K, Neisa A, et al. Estimating the re-identification risk of clinical data sets. *BMC medical informatics and decision making* 2012; 12: 1-15.
16. El Emam K, Arbuckle L, Koru G, et al. De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. *J Med Internet Res* 2012; 14: e33. DOI: 10.2196/jmir.2001.
17. Rodriguez A, Lewis S, Eldridge S, et al. What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol The University of Edinburgh 01 Dec 2020 2020.
18. El Emam K. Chapter 16 - Measuring the Probability of Re-Identification. *Guide to the de-identification of personal health information*. CRC Press, 2013.
19. Rodriguez A, Tuck C, Dozier MF, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. University of Edinburgh, 2020.
20. SAS Institute Inc. SAS 9.4 [Computer software] TS level 1M4, Copyright © 2016 SAS Institute Inc. Cary, NC, USA: SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration., 2013.
21. CEO Roundtable on Cancer Inc. Project Data Sphere, https://www.projectdatasphere.org/ (2020).
22. The Yale University. Yale University Open Data Access (YODA) Project, http://yoda.yale.edu/ (2020, 2020).
23. The University of Edinburgh. Working with sensitive data https://www.ed.ac.uk/information-services/research-support/research-data-service/during/sensitive-data (2020, accessed 30 Oct 2020 2020).
24. The University of Edinburgh. Use University services, https://www.ed.ac.uk/infosec/information-protection-policies/procedures-guidance/use-university-services (2020, accessed 30 Oct 2020 2020).
25. The University of Edinburgh. Data - Data Services, https://www.ed.ac.uk/information-services/research-support/research-computing/ecdf/data (2020, accessed 30 Oct 2020 2020).
26. Microsoft. MS Excel. 2016.
27. Hrynaszkiewicz I, Norton ML, Vickers AJ, et al. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials* 2010; 11. DOI: 10.1136/bmj.c181. 10.1186/1745-6215-11-9.

449 28. El Emam K and Dankar FK. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*
450 2008; 15: 627-637. DOI: 10.1197/jamia.M2716.
451 29. Bogle B and Erickson J. A Moment-Matching Approach for Generating Synthetic Data in SAS®. 2017.
452 30. IOM (Institute of Medicine). *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington (DC): National
453 Academies Press, 2015.
454 31. StataCorp. Stata Statistical Software. College Station, TX: StataCorp LLC.2017.
455 32. National Heart Lung and Blood Institute (NHLBI). Instructions for Preparing Clinical Research Study Datasets for Submission to
456 the NHLBI, https://www.nhlbi.nih.gov/grants-and-training/policies-and-guidelines/guidelines-for-preparing-clinical-study-data-sets-for-
457 submission-to-the-nhlbi-data-repository (2023, accessed 01 Mar 2024 2024).
458 33. Weinreich SS, Mangon R, Sikkens J, et al. Orphanet: a European database for rare diseases. *Nederlands tijdschrift voor*
459 *geneeskunde* 2008; 152: 518-519.
460 34. Mowat C, Arnott I, Cahill A, et al. Mercaptopurine versus placebo to prevent recurrence of Crohn's disease after surgical
461 resection (TOPPIC): a multicentre, double-blind, randomised controlled trial. *The lancet Gastroenterology & hepatology* 2016; 1: 273-282.
462 DOI: 10.1016/s2468-1253(16)30078-4.
463 35. Group ISTC. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19
464 435 patients with acute ischaemic stroke. *The Lancet* 1997; 349: 1569-1581.
465 36. Salman RA-S, Dennis M, Sandercock P, et al. Effects of antiplatelet therapy after stroke due to intracerebral haemorrhage
466 (RESTART): a randomised, open-label trial. *The Lancet* 2019; 393: 2613-2623.
467 37. European Parliament - Council of the European Union. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT
468 AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the
469 free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European
470 Union, 2016, p. 1-78.
471 38. The National Archives. United Kingdom General Data Protection Regulation (UK GDPR) - Regulation (EU) 2016/679 of the
472 European Parliament and of the Council - Regulations originating from the EU. 2016.
473 39. Tudur Smith C, Nevitt S, Appelbe D, et al. Resource implications of preparing individual participant data from a clinical trial to
474 share with external researchers. *Trials* 2017; 18: 319. DOI: 10.1186/s13063-017-2067-4.
475 40. Naudet F, Sakarovitch C, Janiaud P, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical
476 journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. *BMJ* 2018; 360: k400. DOI:
477 10.1136/bmj.k400.
478 41. Clinical Data Interchange Standards Consortium. Study Data Tabulation Model (SDTM),
479 https://www.cdisc.org/standards/foundational/sdtm (accessed 22 Mar 2024 2024).
480 42. Clinical Data Interchange Standards Consortium. Analysis Data Model (ADaM),
481 https://www.cdisc.org/standards/foundational/adam (accessed 22 Mar 2024 2024).
482 43. Griffiths E, Greci C, Kotrotsios Y, et al. Handbook on statistical disclosure control for outputs. *Safe Data Access Professionals*
483 *Working Group* 2019.
484 44. Sweeney L. k-Anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based*
485 *Systems* 2002. DOI: 10.1142/S0218488502001648.
486 45. El Emam K, Middleton G and Arbuckle L. An implementation guide for data anonymization. Bloomington, IN: Trafford
487 Publishing, 2014.
488 46. Keerie C, Tuck C, Milne G, et al. Data sharing in clinical trials - practical guidance on anonymising trial datasets. *Trials* 2018;
489 19: 25. DOI: 10.1186/s13063-017-2382-9.
490 47. Pilgram L, Meurers T, Malin B, et al. The Costs of Anonymization: Case Study Using Clinical Data. *Journal of Medical Internet*
491 *Research* 2024; 26: e49445.
492 48. Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity; Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge*
493 *Discovery from Data* 2007; 1: 3-es. DOI: 10.1145/1217299.1217302.
494 49. Li N, Li T and Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *2007 IEEE 23rd International*
495 *Conference on Data Engineering, Istanbul* 2007; 2: 106-115. DOI: 10.1109/ICDE.2007.367856.
496 50. Dwork C. Differential Privacy: A Survey of Results. *Microsoft Research* 2008: 1-19. DOI: 10.1007/978-3-540-79228-4_38.
497 51. Bull S, Roberts N and Parker M. Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public
498 Health Research: A Systematic Scoping Review. *Journal of empirical research on human research ethics : JERHRE* 2015; 10: 225-238.
499 DOI: 10.1177/1556264615594767.
500 52. Accenture. The Ethics of Data Sharing. 2016.
501 53. Ali J, Califf R and Sugarman J. Anticipated Ethics and Regulatory Challenges in PCORnet: The National Patient-Centered
502 Clinical Research Network. *ACCOUNTABILITY IN RESEARCH-POLICIES AND QUALITY ASSURANCE* 2016; 23: 79-96. DOI:
503 10.1080/08989621.2015.1023951.
504 54. Rocher L, Hendrickx JM and De Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using
505 generative models. *Nature communications* 2019; 10: 1-9.
506 55. Henriksen-Bulmer J and Jeary S. Re-identification attacks—A systematic literature review. *International Journal of Information*
507 *Management* 2016; 36: 1184-1192. DOI: 10.1016/j.ijinfomgt.2016.08.002.
508 56. El Emam K, Jonker E, Arbuckle L, et al. A Systematic Review of Re-Identification Attacks on Health Data. *Plos one* 2011. DOI:
509 10.1371/journal.pone.0028071.
510 57. Janmey V and Elkin PL. Re-identification risk in HIPAA de-identified datasets: The MVA attack. In: *AMIA Annual Symposium*
511 *Proceedings* 2018, p.1329. American Medical Informatics Association.
512 58. Templ M, Meindl B, Kowarik A, et al. Package sdcMicro. 5.5.1 ed. 2020.
513 59. Templ M, Kowarik A and Meindl B. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of*
514 *Statistical Software* 2015; 67: 1-36. DOI: 10.18637/jss.v067.i04.
515 60. Templ M, Meindl B and Kowarik A. Package sdcMicro. 5.6.0 ed. 2021.

516 61. London School of Hygiene & Tropical Medicine. LSHTM Data Compass, https://datacompass.lshtm.ac.uk (2020).
517 62. Fuhr DC, Weobong B, Lazarus A, et al. Delivering the Thinking Healthy Programme for perinatal depression through peers: an
518 individually randomised controlled trial in India. *The Lancet Psychiatry* 2019; 6: 115-127.
519 63. Sikander S, Ahmad I, Atif N, et al. Delivering the Thinking Healthy Programme for perinatal depression through volunteer peers:
520 a cluster randomised controlled trial in Pakistan. *The Lancet Psychiatry* 2019; 6: 128-139.
521 64. Dhalla K, Cousens S, Bowman R, et al. Is beta radiation better than 5 flurouracil as an adjunct for trabeculectomy surgery when
522 combined with cataract surgery? A randomised controlled trial. *PLoS One* 2016; 11: e0161674.
523 65. Clinical Trials Unit London School of Hygiene & Tropical Medicine. The FreeBIRD Bank of Injury and Emergency Research
524 Data, https://freebird.lshtm.ac.uk/ (2020).
525 66. Collaborators CT. Effect of intravenous corticosteroids on death within 14 days in 10 008 adults with clinically significant head
526 injury (MRC CRASH trial): randomised placebo-controlled trial. *The Lancet* 2004; 364: 1321-1328.
527 67. collaborators C-t. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with
528 significant haemorrhage (CRASH-2): A randomised, placebo-controlled trial. *the Lancet* 2010; 376: 376: 323-332. DOI:
529 https://doi.org/10.1016/S0140-6736(10)60835-5.
530 68. Shakur H, Roberts I, Fawole B, et al. Effect of early tranexamic acid administration on mortality, hysterectomy, and other
531 morbidities in women with post-partum haemorrhage (WOMAN): an international, randomised, double-blind, placebo-controlled trial. *The*
532 *Lancet* 2017; 389: 2105-2116.
533 69. Shakur-Still H, Roberts I, Fawole B, et al. Effect of tranexamic acid on coagulation and fibrinolysis in women with postpartum
534 haemorrhage (WOMAN-ETAC): a single-centre, randomised, double-blind, placebo-controlled trial. *Wellcome open research* 2018; 3.
535 70. Grassin-Delyle S, Semeraro M, Lamy E, et al. Pharmacokinetics of tranexamic acid after intravenous, intramuscular, and oral
536 routes: a prospective, randomised, crossover trial in healthy volunteers. *British Journal of Anaesthesia* 2022; 128: 465-472.
537 71. The University of Edinburgh. Edinburgh DataShare, https://datashare.ed.ac.uk/ (2020).
538 72. Lewis SC, Bhattacharya S, Wu O, et al. Gabapentin for the management of chronic pelvic pain in women (GaPP1): a pilot
539 randomised controlled trial. *PloS one* 2016; 11: e0153037.
540 73. Group I-C. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute
541 ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *The Lancet* 2012; 379: 2352-2363.
542 74. Clinical Study Data Request (CSDR). Clinical Study Data Request, https://clinicalstudydatarequest.com/ (2020, 2020).
543 75. Hayden FG, Osterhaus AD, Treanor JJ, et al. Efficacy and safety of the neuraminidase inhibitor zanamivir in the treatment of
544 influenzavirus infections. *New England Journal of Medicine* 1997; 337: 874-880.
545 76. Singh D, Worsley S, Zhu C-Q, et al. Umeclidinium/vilanterol versus fluticasone propionate/salmeterol in COPD: a randomised
546 trial. *BMC Pulmonary Medicine* 2015; 15: 1-12.
547 77. Lugogo N, Domingo C, Chanez P, et al. Long-term efficacy and safety of mepolizumab in patients with severe eosinophilic
548 asthma: a multi-center, open-label, phase IIIb study. *Clinical therapeutics* 2016; 38: 2058-2070. e2051.
549 78. Arzimanoglou A, Ferreira J, Satlin A, et al. Evaluation of long-term safety, tolerability, and behavioral outcomes with adjunctive
550 rufinamide in pediatric patients (≥ 1 to< 4 years old) with Lennox-Gastaut syndrome: final results from randomized study 303. *European*
551 *Journal of Paediatric Neurology* 2019; 23: 126-135.
552 79. Grossberg GT, Farlow MR, Meng X, et al. Evaluating high-dose rivastigmine patch in severe Alzheimer's disease: analyses with
553 concomitant memantine usage as a factor. *Current Alzheimer Research* 2015; 12: 53-60.
554 80. Dryad. Data Dryad, https://datadryad.org/ (accessed 2020).
555 81. Darlow B, Stanley J, Dean S, et al. The Fear Reduction Exercised Early (FREE) approach to management of low back pain in
556 general practice: a pragmatic cluster-randomised controlled trial. *PLoS medicine* 2019; 16: e1002897.
557 82. Zee K-Y, Chan PS, Ho JCS, et al. Adjunctive use of modified Yunu-Jian in the non-surgical treatment of male smokers with
558 chronic periodontitis: a randomized double-blind, placebo-controlled clinical trial. *Chinese Medicine* 2016; 11: 1-13.
559 83. Christopher PP, Appelbaum PS, Truong D, et al. Reducing therapeutic misconception: A randomized intervention trial in
560 hypothetical clinical trials. *PLoS One* 2017; 12: e0184224.
561 84. Mahmud R, Rahman MM, Alam I, et al. Ivermectin in combination with doxycycline for treating COVID-19 symptoms: a
562 randomized trial. *Journal of International Medical Research* 2021; 49: 03000605211013550.
563 85. Fizazi K, Tran N, Fein L, et al. Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. *New England*
564 *Journal of Medicine* 2017; 377: 352-360.
565 86. Raghu G, Million-Rousseau R, Morganti A, et al. Macitentan for the treatment of idiopathic pulmonary fibrosis: the randomised
566 controlled MUSIC trial. *European Respiratory Journal* 2013; 42: 1622-1632.
567 87. Huffstutter JE, Kafka S, Brent LH, et al. Clinical response to golimumab in rheumatoid arthritis patients who were receiving
568 etanercept or adalimumab: results of a multicenter active treatment study. *Current Medical Research and Opinion* 2017; 33: 657-666.
569 88. Leyland-Jones B, Semiglazov V, Pawlicki M, et al. Maintaining normal hemoglobin levels with epoetin alfa in mainly nonanemic
570 patients with metastatic breast cancer receiving first-line chemotherapy: a survival study. *Journal of Clinical Oncology* 2005; 23: 5960-
571 5972.
572 89. O'Connell MJ, Colangelo LH, Beart RW, et al. Capecitabine and oxaliplatin in the preoperative multimodality treatment of rectal
573 cancer: surgical end points from National Surgical Adjuvant Breast and Bowel Project trial R-04. *Journal of clinical oncology* 2014; 32:
574 1927.
575 90. de Alarcon PA, Matthay KK, London WB, et al. Intravenous immunoglobulin with prednisone and risk-adapted chemotherapy for
576 children with opsoclonus myoclonus ataxia syndrome associated with neuroblastoma (ANBL00P3): a randomised, open-label, phase 3
577 trial. *The Lancet Child & Adolescent Health* 2018; 2: 25-34.
578 91. Sparano JA, Gray RJ, Makower DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New*
579 *England Journal of Medicine* 2018; 379: 111-121.
580 92. Baggstrom MQ, Socinski MA, Wang XF, et al. Maintenance sunitinib following initial platinum-based combination
581 chemotherapy in advanced-stage IIIB/IV non–small cell lung cancer: a randomized, double-blind, placebo-controlled phase III study—
582 CALGB 30607 (Alliance). *Journal of Thoracic Oncology* 2017; 12: 843-849.

583 93. Eiermann W, Pienkowski T, Crown J, et al. Phase III study of doxorubicin/cyclophosphamide with concomitant versus sequential
584 docetaxel as adjuvant treatment in patients with human epidermal growth factor receptor 2-normal, node-positive breast cancer: BCIRG-
585 005 trial. *J Clin Oncol* 2011; 29: 3877-3884.
586 94. Pfizer. *Online report for Open Label, Randomised Multicentre Phase III Study Of Irinotecan Hydrochloride (Campto*
587 *(Registered)) And Cisplatin Versus Etoposide And Cisplatin In Chemotherapy Naive Patients With Extensive Disease - Small Cell Lung*
588 *Cancer*. 2010.
589 95. Poulin-Costello M, Azoulay L, Van Cutsem E, et al. An analysis of the treatment effect of panitumumab on overall survival from
590 a phase 3, randomized, controlled, multicenter trial (20020408) in patients with chemotherapy refractory metastatic colorectal cancer.
591 *Targeted oncology* 2013; 8: 127-136.
592 96. Fizazi K, Higano CS, Nelson JB, et al. Phase III, randomized, placebo-controlled study of docetaxel in combination with
593 zibotentan in patients with metastatic castration-resistant prostate cancer. *Journal of Clinical Oncology* 2013; 31: 1740-1747.
594 97. Michaelson MD, Oudard S, Ou Y-C, et al. Randomized, placebo-controlled, phase III trial of sunitinib plus prednisone versus
595 prednisone alone in progressive, metastatic, castration-resistant prostate cancer. *J Clin Oncol* 2014; 32: 76-82.
596 98. The National Heart LaBIN. BioLINCC, https://biolincc.nhlbi.nih.gov/ (2020).
597 99. Network IPFCR. Randomized trial of acetylcysteine in idiopathic pulmonary fibrosis. *New England Journal of Medicine* 2014;
598 370: 2093-2101.
599 100. Group TtRAtPS. Leukocyte reduction and ultraviolet B irradiation of platelets to prevent alloimmunization and refractoriness to
600 platelet transfusions. *New England Journal of Medicine* 1997; 337: 1861-1870.
601 101. Krishnan A, Pasquini MC, Logan B, et al. Autologous haemopoietic stem-cell transplantation followed by allogeneic or
602 autologous haemopoietic stem-cell transplantation in patients with multiple myeloma (BMT CTN 0102): a phase 3 biological assignment
603 trial. *The lancet oncology* 2011; 12: 1195-1203.
604 102. Raghu G, Pellegrini CA, Yow E, et al. Laparoscopic anti-reflux surgery for the treatment of idiopathic pulmonary fibrosis
605 (WRAP-IPF): a multicentre, randomised, controlled phase 2 trial. *The Lancet Respiratory Medicine* 2018; 6: 707-714.
606 103. Lytle LA, Laska MN, Linde JA, et al. Weight-gain reduction among 2-year college students: the CHOICES RCT. *American*
607 *Journal of Preventive Medicine* 2017; 52: 183-191.
608 104. Hochman JS, Lamas GA, Buller CE, et al. Coronary intervention for persistent occlusion after myocardial infarction. *New*
609 *England Journal of Medicine* 2006; 355: 2395-2407.
610 105. The National Institute of Mental Health. The NIMH Data Archive (NDA), https://nda.nih.gov/ (2020).
611 106. Sachs GS, Nierenberg AA, Calabrese JR, et al. Effectiveness of adjunctive antidepressant treatment for bipolar depression. *New*
612 *England Journal of Medicine* 2007; 356: 1711-1722.
613 107. Kerwin ML. Using SMART Treatment Design to Evaluate Applied Behavior Analysis Interventions on Communication in
614 Preschool Children with Autism.
615 108. Kelly DL, Demyanovich HK, Rodriguez KM, et al. Randomized controlled trial of a gluten-free diet in patients with
616 schizophrenia positive for antigliadin antibodies (AGA IgG): a pilot feasibility study. *Journal of Psychiatry and Neuroscience* 2019; 44:
617 269-276.
618 109. Sikich L, Kolevzon A, King BH, et al. Intranasal oxytocin in children and adolescents with autism spectrum disorder. *New*
619 *England Journal of Medicine* 2021; 385: 1462-1473.
620 110. Group HDTS and Group HDTS. Effect of Hypericum perforatum (St John's wort) in major depressive disorder: a randomized
621 controlled trial. *Jama* 2002; 287: 1807-1814.
622 111. Vivli Center for Global Clinical Research Data. Vivli, a global data-sharing and analytics platform. , https://vivli.org/ (2020,
623 2020).
624 112. Langley JM, Carmona Martinez A, Chatterjee A, et al. Immunogenicity and safety of an inactivated quadrivalent influenza
625 vaccine candidate: a phase III randomized controlled trial in children. *The Journal of infectious diseases* 2013; 208: 544-553.
626 113. Oliver AJ, Covar RA, Goldfrad CH, et al. Randomised trial of once-daily vilanterol in children with asthma on inhaled
627 corticosteroid therapy. *Respiratory Research* 2016; 17: 1-11.
628 114. Calverley PM, Anderson JA, Brook RD, et al. Fluticasone furoate, vilanterol, and lung function decline in patients with moderate
629 chronic obstructive pulmonary disease and heightened cardiovascular risk. *American Journal of Respiratory and Critical Care Medicine*
630 2018; 197: 47-55.
631 115. GlaxoSmithKline group of companies. *A Randomised, Double-blind, Placebo-controlled, Incomplete Block, 4-period Crossover,*
632 *Study to Investigate the Effects of 5-day Repeat Inhaled Doses of Fluticasone Propionate (BID, 50-2000 mcg) on Airway Responsiveness to*
633 *Adenosine 5-monophosphate (AMP) Challenge When Delivered After the Last Dose in Mild Asthmatic Subjects. (GSK 04_SIG103337).*
634 *Clinical Summary Report* 20 October 2006 2006.
635 116. Kim JH, Lee SK, Loesch C, et al. Comparison of levetiracetam and oxcarbazepine monotherapy among Korean patients with
636 newly diagnosed focal epilepsy: A long-term, randomized, open-label trial. *Epilepsia* 2017; 58: e70-e74.
637 117. UK Data Service. UK Data Service: data Catalogue, https://beta.ukdataservice.ac.uk/datacatalogue/studies (accessed 2020).
638 118. Csipke E, Shafayat A, Sprange K, et al. Promoting independence in dementia (PRIDE): A feasibility randomized controlled trial.
639 *Clinical Interventions in Aging* 2021: 363-378.
640 119. Yiend J, Lam CL, Schmidt N, et al. Cognitive bias modification for paranoia (CBM-pa): a randomised controlled feasibility study
641 in patients with distressing paranoid beliefs. *Psychological medicine* 2023; 53: 4614-4626.
642 120. Bracken H, Mundle S, Faragher B, et al. Induction of labour in pre-eclamptic women: a randomised trial comparing the Foley
643 balloon catheter with oral misoprostol. *BMC pregnancy and childbirth* 2014; 14: 1-5.
644 121. McEwan K, Richardson M, Sheffield D, et al. A smartphone app for improving mental health through connecting with urban
645 nature. *International journal of environmental research and public health* 2019; 16: 3373.
646 122. Murphy AW, Cupples M, Smith S, et al. Effect of tailored practice and patient care plans on secondary prevention of heart
647 disease in general practice: cluster randomised controlled trial. *Bmj* 2009; 339.
648 123. Intersect Australia Limited - Queensland Cyber Infrastructure Foundation Ltd. Australian National Medical Research Data
649 Storage Facility, https://med.data.edu.au/find-data/ (accessed 2020).

650 124. Institute DCR. SOAR DATA™, https://dcri.org/our-approach/data-sharing/soar-data (2020).
651 125. PLOS is a nonprofit 501(c)(3) corporation. PLOS ONE: An inclusive journal community working together to advance science by
652 making all rigorous research accessible without barriers, https://journals.plos.org/plosone/search.
653 126. Anjara SG, Bonetto C, Ganguli P, et al. Can General Practitioners manage mental disorders in primary care? A partially
654 randomised, pragmatic, cluster trial. *PLoS One* 2019; 14: e0224724.
655 127. Vijitpavan A, Kittikunakorn N and Komonhirun R. Comparison between intrathecal morphine and intravenous patient control
656 analgesia for pain control after video-assisted thoracoscopic surgery: A pilot randomized controlled study. *Plos one* 2022; 17: e0266324.
657 128. Chowdhury F, Shahid ASMSB, Tabassum M, et al. Vitamin D supplementation among Bangladeshi children under-five years of
658 age hospitalised for severe pneumonia: A randomised placebo controlled trial. *Plos one* 2021; 16: e0246460.
659 129. Weinberg L, Ianno D, Churilov L, et al. Restrictive intraoperative fluid optimisation algorithm improves outcomes in patients
660 undergoing pancreaticoduodenectomy: a prospective multicentre randomized controlled trial. *PLoS One* 2017; 12: e0183313.
661 130. Choi W, Kim JC, Kim WS, et al. Clinical effect of antioxidant glasses containing extracts of medicinal plants in patients with dry
662 eye disease: a multi-center, prospective, randomized, double-blind, placebo-controlled trial. *PLoS One* 2015; 10: e0139761.
663 131. Iannotti L, Dulience SJ-L, Joseph S, et al. Fortified snack reduced anemia in rural school-aged children of Haiti: a cluster-
664 randomized, controlled trial. *PloS one* 2016; 11: e0168121.
665 132. BMJ Publishing Group Ltd. BMJ is a global healthcare knowledge provider with a vision for a healthier world. We share
666 knowledge and expertise to improve healthcare outcomes., https://www.bmj.com/search/advanced.
667 133. Ebbeling CB, Feldman HA, Klein GL, et al. Effects of a low carbohydrate diet on energy expenditure during weight loss
668 maintenance: randomized trial. *bmj* 2018; 363.
669 134. Nave AH, Rackoll T, Grittner U, et al. Physical Fitness Training in Patients with Subacute Stroke (PHYS-STROKE): multicentre,
670 randomised controlled, endpoint blinded trial. *Bmj* 2019; 366.
671 135. Costa ML, Achten J, Ooms A, et al. Surgical fixation with K-wires versus casting in adults with fracture of distal radius:
672 DRAFFT2 multicentre randomised clinical trial. *bmj* 2022; 376.
673 136. Harvard University. Harvard Dataverse Repository. Deposit and share your data. Get academic credit. Harvard Dataverse is a
674 repository for research data. Deposit data and code here., https://dataverse.harvard.edu/.
675 137. Gehani M, Kapur S, Madhuri SD, et al. Effectiveness of antenatal screening of asymptomatic bacteriuria in reduction of
676 prematurity and low birth weight: Evaluating a point-of-care rapid test in a pragmatic randomized controlled study. *EClinicalMedicine*
677 2021; 33.
678 138. Elson L, Randu K, Feldmeier H, et al. Efficacy of a mixture of neem seed oil (Azadirachta indica) and coconut oil (Cocos
679 nucifera) for topical treatment of tungiasis. A randomized controlled, proof-of-principle study. *PLoS Neglected Tropical Diseases* 2019;
680 13: e0007822.
681 139. Semrau KE, Hirschhorn LR, Marx Delaney M, et al. Outcomes of a coaching-based WHO safe childbirth checklist program in
682 India. *New England Journal of Medicine* 2017; 377: 2313-2324.
683 140. Jayawardane M, Piyadigama I and Chandradeva U. Will a preoperative theatre visit reduce anxiety? A randomised controlled
684 trial. *Journal of Obstetrics and Gynaecology* 2022; 42: 1498-1503.
685 141. Stitely ML, Harlow K and MacKenzie E. Oral riboflavin to assess ureteral patency during cystoscopy: a randomized clinical trial.
686 *Obstetrics & Gynecology* 2019; 133: 301-307.
687 142. Antibacterial Resistance Leadership Group (ARLG) ARLG studies, https://arlg.org/summary-of-results/.
688 143. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). NIDDK Central Repository,
689 https://repository.niddk.nih.gov/studies/dpp/.
690

4.3 Extended methods, results and discussion

In this part of chapter 4, I present important broader components of the research that, while significant, were not included in the proposed manuscript as they were not part of the study protocol. The protocol focused solely on addressing the second and third objectives of this PhD: “To investigate whether individual participants could potentially be at risk of being re-identified from a range of datasets that have been anonymised and made available for sharing” and “To identify factors that could increase the risk of re-identification of an anonymised clinical trial dataset.”

First, the protocol for calculating re-identification risk scores had to be carefully crafted to convey that our aim was not to identify individuals but to assess any potential risk present in the datasets. For example, the protocol's title was changed from “What is the risk of re-identification of individuals within publicly available anonymised/de-identified clinical trial datasets? A study protocol” to “What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol”. I successfully explained that the re-identification risk calculations did not pose a risk to individuals in the anonymised datasets, as evidenced by the high success rate of obtaining datasets.

Second, during the protocol revision stage, I discovered several sources describing how to calculate re-identification risk scores. These included methodologies by the Comprehensive R Archive Network (CRAN) via its “Statistical Disclosure Control for Micro-Data Using the R Package sdcMICRO” (Templ, Kowarik et al. 2015, Templ, Meindl et al. 2021), “Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk”(IOM (Institute of Medicine) 2015), and “The Anonymisation Decision-Making Framework (ADF)”(Elliot 2016, Elliot, Mackey et al. 2020). Initially, these methodologies were included in the draft protocol. However, they are general and not particularly tailored to clinical trial datasets. Upon closer inspection, I decided to use only the methodology by El Emam for the final

protocol, as it is the most comprehensive, well-documented, and encompasses all the previously mentioned methodologies. I also tested this methodology with the ECTU non-anonymised datasets for the study “Fractures and Bisphosphonates: A Double-Blind, Randomised Controlled Trial on the Effect of Alendronic Acid on Healing and Clinical Outcomes of Wrist Fractures”, the FaB study (Duckworth, McQueen et al. 2019), to ensure that the methodology could be implemented on clinical trials datasets. At this stage, the re-identification report consisted of two parts. The first part presented a cross-tabulation of the indirect identifiers, a necessary step for calculating re-identification risk scores. The second part presented the calculated risk scores. I realised that the first part could potentially be highly disclosive, so for the final report, I requested the statistical software to suppress its printing. I also recognised that non-manipulation of the datasets should be a mandatory protocol requirement and that there was very little value in exploring combinations of available indirect identifiers, as datasets should be analysed as they are. The results of the methodology test are presented in Appendix 5 part 1.

Thirdly, I underestimated the effort and time required to sign data-sharing agreements between my institution and third parties. This was mainly due to two factors: the COVID-19 pandemic and personnel shortages at Edinburgh Research Office (ERO) at the University of Edinburgh. Despite these challenges, ERO became an invaluable partner and successfully signed all the necessary data-sharing agreements. Additional materials not submitted for publication include the compiled file with all datasets’ metadata in Appendix 4 and each dataset's individual re-identification risk score report in Appendix 5 part 2. I consider these materials sensitive and confidential, and therefore not suitable for the public domain. My goal is to further knowledge in the area of anonymisation of clinical trials datasets. These documents have the potential to expose vulnerabilities in the analysed datasets, and the aim is not to name and shame

data holder/owners but to increase understanding of how we can further protect anonymised datasets.

I am also attaching the relevant SAS code in Appendix 6, which I am offering to share at the end of the manuscript with those who contact us. This has been deliberately left out of the submitted materials as I hope to engage with interested parties, further enhancing my knowledge of the topic through their needs and creating opportunities for future collaborations.

Preliminary results from available datasets as of the 30th June 2022 (44 datasets) were selected as an oral presentation at the 6th International Clinical Trials Methodology Conference (ICTMC) in October 2022 in Harrogate, UK, after successfully submitting an abstract to this conference. This talk concentrated on the process to obtaining the datasets, the characteristics of the datasets, and the gaps to fully executing the protocol. I then presented the final results of the 70 analysed datasets at Statisticians in the Pharmaceutical Industry (PSI) conference in June 2024 in Amsterdam, Netherlands. In both instances, the topic was well received by the audience. Finally, I am scheduled to give an oral presentation at the 7th ICTMC in October 2024 in Edinburgh, UK. This is all evidence that there is an appetite in the clinical trials community to improve practices to further protect clinical trials datasets.

4.4 Conclusion

In this part of the thesis, I performed an exploratory analysis on individual patient data (IPD) datasets shared for secondary research from 70 clinical trials. I calculated the re-identification risk scores using the methodologies proposed by El Emam (El Emam 2013) to determine if these scores could be applied to clinical trials and, if so, whether they serve any purpose in the anonymisation process of such datasets.

The results presented in this chapter highlight a significant tendency among researchers to share clinical trial datasets, which often include sensitive personal information. Although re-identification risk scores for these datasets might seem overly high (especially when compared to larger electronic health record datasets in the literature), these high scores do not necessarily indicate a real re-identification threat. The scores do not reflect the actual vulnerability of datasets to re-identification attacks in the real world, as they are influenced by factors such as an attacker's motivations, resources, and potential gains. Instead, these scores provide important insights into the level of detail in clinical trial datasets, aiding decisions about data release for secondary use.

I showed that calculating re-identification risk scores is simple and practical. The number and type of identifiers (continuous or discrete) are crucial in managing risk scores. More identifiers increase granularity, with continuous identifiers adding more detail than discrete ones. While risk scores alone cannot ensure sufficient anonymisation, they assist in fine-tuning the process of anonymising clinical trial datasets.

Clinical trial researchers could compare the re-identification risk scores from their anonymised datasets to the scores calculated in this research to inform their anonymisation processes. Researchers are also advised to use the outlined procedure in the proposed manuscript to guide their anonymisation efforts before data release. The proposed method is both cost-effective and practical. Despite its simplicity, it is highly informative and serves as an effective interim measure, particularly when resources are limited. In the next chapter, I will explore whether re-identification risk scores (commonly used in health records) are known and used in the UK clinical trials community.

Chapter 5 UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation

5.1 Introduction

In Chapter 3, I gathered a collection of recommendations and guidance on how to anonymise clinical trial datasets. In Chapter 4, I discovered that re-identification risk scores, commonly used for health records, could inform the process of sharing anonymised clinical trial datasets for secondary research. With these insights in hand, the natural next questions were whether my peers in clinical trials are using recommendations and guidance to anonymise their datasets for secondary purposes, their experiences so far with the process of anonymising clinical trial data, and their general experiences with data sharing. Additionally, I wanted to explore whether they were aware of and using re-identification risk scores to inform the process of dataset release. This relates to the fourth objective of my PhD: To explore researchers' views and experiences regarding the sharing of clinical trial data.

To address these questions, I decided to execute a cross-sectional survey to gather those opinions, as cross-sectional surveys are an efficient and cost-effective way to collect data from a representative group at one specific point in time (Creswell 2008).

Therefore, I developed a protocol titled "What are the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets? A study protocol" along with an "online exploratory cross-sectional descriptive survey". I used the protocol and the survey to apply for ethical research approval on the 4th May 2022 (additional files 1 and 2 in Appendix 7) and obtained a favourable ethical opinion (22-EMREC-027) on the 1st June 2022 (additional file 7 in Appendix 7). After implementing the survey and completing all protocol activities, I compiled a manuscript for the publications of the

results. I meticulously addressed all comments from my co-authors and then submitted the manuscript to the peer-reviewed Clinical Trials SAGE Journal on the 19th May 2023. I addressed all the peer reviewers' comments, the manuscript was accepted for publication on the 5th April 2024. The results of the cross-sectional survey were published as a peer-reviewed paper in the Clinical Trials SAGE Journal on the 19th June 2024 and is openly licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Our article, titled "A survey on UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trial datasets" (Rodriguez, Lewis et al. 2024) , is available at https://doi.org/10.1177/17407745241259086 and is presented in the next item of this chapter. All additional published materials related to this article are included in Appendix 7.

5.2 Published article

Article

CLINICAL TRIALS

Clinical Trials
1–13

DOI: 10.1177/17407745241259086
journals.sagepub.com/home/ctj
Sage

A survey on UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trial datasets

Aryelly Rodriguez[1], Steff C Lewis[1], Sandra Eldridge[2], Tracy Jackson[3] and Christopher J Weir[1]

Abstract

Background: There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. However, there is no standardised set of recommendations on how to anonymise and prepare clinical trial datasets for sharing, while an ever-increasing number of anonymised datasets are becoming available for secondary research. Our aim was to explore the current views and experiences of researchers in the United Kingdom about de-identification, anonymisation, release methods and re-identification risk estimation for clinical trial datasets.

Methods: We used an online exploratory cross-sectional descriptive survey that consisted of both open-ended and closed questions.

Results: We had 38 responses to invitation from June 2022 to October 2022. However, 35 participants (92%) used internal documentation and published guidance to de-identify/anonymise clinical trial datasets. De-identification, followed by anonymisation and then fulfilling data holders' requirements before access was granted (controlled access), was the most common process for releasing the datasets as reported by 18 (47%) participants. However, 11 participants (29%) had previous knowledge of re-identification risk estimation, but they did not use any of the methodologies. Experiences in the process of de-identifying/anonymising the datasets and maintaining such datasets were mostly negative, and the main reported issues were lack of resources, guidance, and training.

Conclusion: The majority of responders reported using documented processes for de-identification and anonymisation. However, our survey results clearly indicate that there are still gaps in the areas of guidance, resources and training to fulfil sharing requests of de-identified/anonymised datasets, and that re-identification risk estimation is an underdeveloped area.

Keywords

Clinical trials, data anonymisation, re-identification, de-identification, data sharing, re-identification risk

Background

There is now a strong drive, particularly from publishers and funders, to encourage the release of relevant anonymised trial data sets.[1] Therefore, data sharing has become an essential activity to disseminate current research, to enable new investigations and to maximise the scientific endeavour.[2,3] Currently, many anonymised datasets are made publicly available for secondary research via open or controlled access.[4–6] Anonymisation of data is complex, and its full

[1]Edinburgh Clinical Trials Unit, Usher Institute, The University of Edinburgh, Edinburgh, UK
[2]Pragmatic Clinical Trials Unit, Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London UK
[3]Asthma UK Centre for Applied Research, Usher Institute, The University of Edinburgh, Edinburgh, UK

Corresponding author:
Aryelly Rodriguez, Edinburgh Clinical Trials Unit, Usher Institute, The University of Edinburgh, Level 2, Nine Edinburgh BioQuarter, 9 Little France Road, Edinburgh EH16 4UX, UK.
Email: aryelly.rodriguez@ed.ac.uk

implementation could mean that the detail necessary to appropriately analyse the data is lost.[7] There is therefore a balance between wanting to de-risk a dataset prior to sharing, against wanting it to be sufficiently detailed to answer valid research questions, and to allow researchers to repeat the original published analysis. In addition, we are currently investigating re-identification risk scores across a range of clinical trial datasets.[8] Re-identification risk scores, as described by in the work by El Emam,[9] are derived from three equations that use information in the anonymised dataset. They are currently used for routinely collected health records and only generate numerical values. These scores do not aim to re-identify individuals in the datasets and could potentially be applicable to clinical trial datasets. Therefore, we explored UK researchers' views regarding their experiences with the creation and release of de-identified/anonymised clinical trial datasets, the generation and use of re-identification risk scores, and their views about wider aspects of re-identification risks. Humphreys et al.[10] covered wider aspects of data sharing in clinical trials while our study focuses on clinical trial datasets that have been anonymised/de-identified.

Why it is important to do this study?

Knowing what is working and what is not regarding the creation and release of de-identified/anonymised clinical trial datasets, and determining if re-identification risk scores are already in use, from UK clinical trials researchers, will help identify areas for improvements and future research.

Objective

This study aimed to explore the clinical trial researchers' views on their experiences with the creation and release of de-identified/anonymised clinical trial datasets, and the generation and use of re-identification risk scores, and the wider aspects of re-identification risks.

Methods

A full protocol (in the supplementary material, Additional File 1) and a survey instrument (in Additional File 2) were finalised on 28 April 2022. A non-personal invitation letter was generated to describe the study to potential participants (Additional File 5), before they fully engaged with the survey. The invitation letter and the first part of the survey emphasised the voluntary nature of participation, the protection and handling of personal data, and confidentiality. Consent was obtained from the respondents to participate in the survey, and they were assured they could stop and dropout at any time during the study without any consequences.

Survey design

The 'checklist of questions for designing a survey study plan' by Creswell and Creswell[11] was followed for the development of this study (see Additional File 3). We used an online exploratory cross-sectional descriptive survey[11,12] that consists of both open-ended and closed questions for data collection. This allowed us to gather information to better describe actual experiences regarding the investigated topic. The open-ended questions were especially important because of the lack of previous reporting on researchers' views and experiences.

The survey was in English. Most of the closed questions had mutually exclusive choices, with a smaller number allowing for multiple answers.[13,14] Where applicable, closed questions, had an 'other' (free text) option added to allow participants to provide an answer that was not available for selection.[13] Five-point response scales were used for questions assessing frequency (*always, often, sometimes, rarely, never*).

The survey was structured in five parts:

- Consent and eligibility check.
- Section 1. Researchers' work background details (current position, years of experience in current position and general place of work)
- Section 2 Researchers' experiences with the creation and release of de-identified/ anonymised clinical trial datasets
- Section 3. Researchers' awareness, knowledge and use regarding the generation of re-identification risk scores as described in the work by El Emam[9]
- Section 4. Researchers' views about wider aspects of re-identification risks

Where applicable, we also provided short explanations of the concepts used in the survey at the beginning of the relevant section to avoid ambiguity and confusion, as follows:

- De-identification refers to the removal of all personal health information and all other indirect identifiers which could lead to the identification of an individual. The most common de-identification methods are HIPPA (US Health Insurance Portability and Accountability Act of 1996) Safe harbour,[15] in which 18 identifiers are removed from the datasets and Hrynaszkiewicz et al.[16] with an enhanced removal of potential identifiers which are commonly present in clinical trial datasets.
- Anonymisation is when a dataset has been de-identified and then subsequent data manipulation/ steps have been taken to further protect the dataset, for example, if a privacy model has been applied (e.g. k-anonymity)[17] or the link with the original

non-anonymised dataset has been destroyed and this action cannot be reversed.
- Data release under controlled access: Datasets that can only be accessed if permission is granted by the data holders via their internal procedures.
- Data release under open access: Datasets that can be accessed without any or minimal restrictions imposed by the data holders.
- Re-identification risk scores[9] are defined as the estimated probabilities of any given individual being re-identified from an anonymised/de-identified dataset. The re-identification risk score depends on the variables available in the dataset, the number of observations in the dataset and on the strategy used to attack the dataset (prosecutor or journalist scenario).
- Prosecutor scenario[9] is when the adversary knows that a target individual (for whom identifiers are known) is in the publicly available dataset (released anonymised and de-identified).
- Journalist scenario[9] is when the adversary sets out to identify any individual from the publicly available dataset just to prove that it can be done using another dataset for 'matching' with the publicly available dataset.

The final version of the survey is presented in Additional File 2 of this study.

The survey was designed to follow the layout presented in Additional File 4. Therefore, a single participant (after the eligibility criteria has been met) answered between 17 and 22 questions out of the proposed 24 questions, as some answers determined the relevance of the next question.

The survey was piloted using a selection of University of Edinburgh personnel with experience in the processes of de-identification/anonymisation, release/maintenance and re-identification risk assessment of clinical trial datasets. It was then finalised and sent to the intended participants.

Study population

Inclusion/Exclusion criteria. Clinical trial researchers based in the United Kingdom with experience in executing/overseeing any of the processes of de-identification/anonymisation, release/maintenance and re-identification risk assessment of clinical trial datasets to prepare them for secondary research.

Sampling and recruitment

There was no formal sample size or stratification of the surveyed researchers as this is an exploratory study. Therefore, we used convenience non-probability sampling[12,18,19] by providing a Microsoft (MS) Form[20] link or quick response (QR) code with an invitation letter (email or printout) (see Additional File 5) to the following:

- All 52 Clinical Trial Units[21] (CTUs, which are the specialised units that design, execute, analyse and publish clinical trials) registered in the UK Clinical Research Collaboration (UKCRC) network.[22] We emailed all UK fully or provisionally registered CTUs (used list is in Additional File 6).
- The data transparency group at the Global Healthcare Data Science Community (Pharmaceutical Users Software Exchange)[23] (Contacted via email, population size unknown).
- Allstat@JISCMAIL.AC.UK, a statistics email discussion list for the UK Education and Research communities[24] (Contacted via email, population size unknown).
- Participants at the Sixth International Clinical Trials Methodology Conference (ICTMC) (3 6 October 2022; Special event) (Contact via leaflet and a QR code in an oral presentation, Population size unknown).

The aim was to obtain as many responses as possible while the main survey was active (around 5 weeks) to maximise the range of experiences. We estimated the population to be heterogeneous, so a minimum of between 12 and 30 surveys was required[18] to reach data saturation[25] and reflect a wide range of views.

Data collection and analysis

This survey did not collect any personal data from the clinical trial researchers, and after extraction, all open questions were carefully checked to make sure their coding did not contain any identifiable information. Only A.R. was able to access all the data. We used MS Forms as it provided a suitable integrated web interface and data collection tool for the survey. The data within MS Forms 'are encrypted both at rest and in transit' and are stored on a European Server, compliant with UK General Data Protection Regulation (UK GPDR).[26,27]

When the active period for the survey ended, the response summary information and the individual responses of the complete surveys were exported from MS Forms directly to A.R.'s DataStore allocation, a secured and password-protected area at the University of Edinburgh, in accordance with their data handling policies.[28–30]

Individual responses were kept until February 2023, then destroyed in accordance with the University of Edinburgh policy for destroying archived research data.[31,32]

Closed questions were analysed using descriptive statistics (counts and percentages) in SAS 9.4.[33] All data were analysed by A.R.

Thematic analysis[34,35] was used to generate themes from the open-ended questions using NVivo® January 2022 (Release 1.6.1).[36] Participants had the freedom to write as much as they wanted and express several opinions for any given topic. The free-text data were initially coded solely by A.R. These themes were then reviewed, refined and finalised on 7 March 2023, through discussion with the multi-disciplinary research team, to ensure valuable perspectives were included and to help reduce the subjectivity of the findings (S.C.L., C.J.W. and T.J.).

The results of this study helped us to understand the views of UK researchers regarding their experiences with the creation and release of de-identified/anonymised clinical trial datasets, the generation and use of re-identification risk scores, and their views about wider aspects of re-identification risks.

Results

The pilot survey was active from 6 June 2022 to 29 August 2022 inclusive, and the main survey was active from 13 September 2022 to 19 October 2022. There were no changes made to the survey between the pilot and main phase. We obtained 52 consented participants in total of which 38 were eligible because they identified as being based in the United Kingdom. No data from the eligible participants were excluded from the analysis. Notably, 32 (84%) participants were associated with a UKCRC-registered CTU. The average time to complete the survey was 18 min and all participants reached the end of the survey, so there is no missing data to report. The most common role was statistician (including senior statistician; 17 (44%)), followed by Director/Senior Manager (6 (16%)) and principal investigator (4 (11%)). However, 27 (71%) participants had at least 6 years of experience in their employed role at the moment they took the survey. Table 1 has more details on the participant characteristics.

The most common involvement with the de-identification/anonymisation datasets was with their creation/generation (27 participants (71%)) and approval (21 participants (55%)) of the release of de-identified/anonymised datasets. Notably, 27 (71%) participants were involved in more than one task. However, 24 (63%) participants had at least 3 years of experience in dealing with de-identification/anonymisation datasets. In addition, 35 (92%) participants used documentation/guidance for de-identification/anonymisation, of which 21 (60% of 35) participants used both internal and external documents/guidance. Moreover, 24 out of 29 (83%) of the internally generated documentation (either implemented or under construction) covered the topic of how to de-identify/anonymise datasets, also 24 out of 29 (83%) covered the releasing of de-identified/anonymised datasets and 11 out of 29 (38%) covered the

Table 1. Participant characteristics and experience and documentation used on the creation/release of de-identified/anonymised datasets.

Parameters	Participants N = 38 n (%)
Place of work	
UKCRC-registered CTU	32 (84)
Other[a]	6 (16)
Employed role	
Director/Senior Manager	6 (16)
Principal Investigator	4 (11)
Researcher (Research fellow/assistant)	2 (5)
Senior Researcher	1 (3)
Statistician	7 (18)
Senior Statistician	10 (26)
Trial Manager/Coordinator	1 (3)
Senior Trial Manager/Coordinator	1 (3)
Other	6 (16)
Years of experience in employed role	
0–2	3 (8)
3–5	8 (21)
6–10	9 (24)
>10	18 (47)
Involvement with the de-identification/anonymisation tasks[b]	
○ Creation/generation of de-identified/anonymised dataset	27 (71)
○ Evaluation/assessment/peer review of de-identified/anonymised dataset	15 (40)
○ Approval of the release of de-identified/anonymised dataset	21 (55)
○ Generation/evaluation/assessment of de-identified/anonymised dataset re-identification risk	10 (26)

(continued)

Table 1. (continued)

Parameters	Participants N = 38 n (%)
○ Uploading/maintenance/distribution of de-identified/anonymised dataset	14 (37)
○ Other[c]	2 (5)
Years of experience in de-identification/anonymisation	
0–2	14 (37)
3–5	12 (32)
6–10	6 (16)
>10	6 (16)
Documents used	
Only internally developed documents/guidance	8 (21)
Only externally sourced documents/guidance	4 (11)
Both internal and external documents/guidance	21 (55)
Other[d]	5 (13)
Topics covered by the internally developed documents/guidance[e]	
• The process of how to de-identify/anonymise clinical trial datasets	
Yes (documents/guidance implemented)	16 (55)
Yes (but document/guidance under construction)	8 (28)
No (this process is not covered)	4 (14)
No response	1 (3)
• The process for releasing de-identified/anonymised clinical trial datasets	
Yes (documents/guidance implemented)	18 (62)
Yes (but document/guidance under construction)	6 (21)
No (this process is not covered)	3 (10)
No response	2 (7)
• The assessment of the re-identification risk of the de-identified/anonymised clinical trial datasets	
Yes (documents/guidance implemented)	4 (14)
Yes (but document/guidance under construction)	7 (24)
No (this process is not covered)	15 (52)
No response	3 (10)
Process use for releasing de-identified/anonymised clinical trial data	
Only de-identification, under controlled access	13 (34)
De-identification followed by anonymisation, under controlled access	18 (47)
Only de-identification, under open access	1 (3)
De-identification followed by anonymisation, under open access	1 (3)
Other[f]	5 (13)

[a]Other = 1 CRO, 1 Medical School, 1 Retired, 2 University, 1 UoE

[b]Participants were allowed to mark multiple types of involvement.

[c]Other = not directly involved

[d]Other = 3 none/not applicable, 1 'NHSD & CAG advise relating to what is classed as identifiable and sensitive' and 1 'Internal and also review the funding body guidance for whoever funded the study'

[e]Only applicable to the 29 participants who answered in the previous question 'Only internally developed documents/guidance' and 'Both internal and external documents/guidance'

[f]Other = 2 not applicable, 1 'De-identification followed by some aspects of anonymisation, under controlled access (pseudonymisation: there is some manipulation, but the data could not be described as anonymised, and the link may not be destroyed)', 1 'Defining these processes is a work in progress for us, but any data releases would be under controlled access', 1 'varies depending on the risk'

assessment of the re-identification risk. De-identification, followed by anonymisation and then fulfilling data holders' requirements before access was granted (controlled access), was the most common process for releasing the datasets with 18 responses (47%). Further detail is presented in Table 1.

Views on the process of de-identifying/anonymising datasets were asked. From thematic analysis of data from 38 participants, we obtained 81 separate opinions on de-identifying/anonymising datasets. However, 63 expressed a negative sentiment and we categorised them as follows: long process (13), lack of advice (13), data constraints and keeping utility (10), lack of resources (9), risky process (8), difficult process (4), non-reversible process (2), not applicable to old datasets (2), forced process (1), and data requestors not willing to pay/wait (1). Moreover, 17 opinions were of a positive nature and were categorised as: straightforward process (9), others do it (3), non-risky process (2), sharing with trusted bona fide researchers (2) and guidance available (1). Table 2 has representative quotes for each category.

Table 2. Opinions about the process of de-identifying/anonymising datasets.

Name of code	No. of comments	Description/representative quote
Opinion process of de-identifying/anonymising	81[a]	In your opinion, how was your experience in the process of de-identifying/anonymising clinical trials datasets? (e.g. what have worked? what have not worked? any concerns? any assurances?)
+ Is data used for secondary research?	1	Is there any value of doing this?
+ Negative	63[b]	
Data constraints and keeping utility	10	Age and gender are often key parts of an analysis, which we can't remove / do the de-identification without rendering the database meaningless
Difficult process	4	This can be a challenging process
Forced process	1	Release of anonymised data was required for legal reasons
Lack of advice	13	It was difficult to get specific advice on how to de-identify and anonymise data. Especially on assessment of anonymised data
Lack of resources	9	we do not have the resources for this–it is hard to fund as it occurs after the end of a study grant. / You need to find someone who can understand the original data collection and databases and who has the time and skill to identify which fields need to be removed
Data requestors not willing to pay/wait	1	It doesn't happen often, and mainly, people don't want the data badly enough to pay for the work or wait.
Not applicable to old datasets	2	We work with older datasets so the technical side of de-identifying data that was collected at a point when none of the current legislation was in place is demanding
Long process	13	Generally time consuming. / It takes a long time to do it properly
Non-reversible process	2	I recall MRC guidance is to generate a new patient reference number and delete code to generate that number. In practice this is impractical as if the requestor needs clarification on an element of the data being shared you have lost the ability to identify that patient within your own 'parent' dataset
Risky process	8	This makes deidentifying/ anonymising an unnerving experience
+ Positive	17[b]	
Non-risky process	2	however in my opinion it is almost never that this puts participants at risk
Sharing with trusted bona fide researchers	2	We share them with bona fide researchers, with good research ideas, and this has not caused us concern
Others do it	3	Others have performed the process
Straight forward process	9	It was fairly straightforward
Guidance available	1	use the ICO guidelines

[a]Only main nodes added (+)
[b]Only children nodes added

Also, opinions regarding the process of releasing datasets were sought. In total, 38 participants provided 59 opinions. However, 24 expressed negative sentiments, such as issues with data sharing agreements (7), lack of resources (5), risky task (5), complex requests (2), paperwork/red tape/regulation (2) and no demand (2). Meanwhile, 29 opinions reflected positive sentiments, such as easy process (7), easy data transfer (6), committees/expert approvals (5), use of data repositories (4), use of controlled environment (3), vetoing researchers (3) and open access (1). Table 3 has further details.

Opinions on the experience of maintaining released de-identified/anonymised datasets were collected: 12 participants expressed that they did not have experience or that it was not a necessary process. From the 26 participants who had experience in this area, we collected 11 negative opinions of this process being a burden (no resources (3), dynamic process (3), issue with contracts (2), bespoke process (1), difficult to keep links (1) and difficult to keep process for access (1)). The only positive opinion was that the maintenance was done for free by the participant's institution (details in Table 4).

In addition, 38 (100%) of the participants did not use any kind of re-identification risk score, this was distributed as follows: 18 (47%) participants had never heard of re-identification risk scores while 9 (24%) and 11 (29%) of the participants have, respectively, 'heard

Table 3. Opinions about the process of release datasets.

Name of code	No. of comments	Description/representative quote
Opinion process of release datasets	59[a]	In your opinion, how was your experience in the process of releasing de-identified/anonymised clinical trials datasets? (e.g. what have worked? what have not worked? any concerns? any assurances?)
+ Done by others	3	Having gone through the de-identifying phase the release was overseen by others
+ Not done this yet	3	Have not done this yet
+ Negative	24[b]	
Complex requests	2	there is considerable conversation regarding the manipulation to end up at the dataset
Process improving	1	The process here is getting better
Issues with data sharing agreements	7	Main issue has been regulatory/contracts rather than data process
Paperwork/red tape/regulation	2	More form filling in / Too much regulation
Lack of resources	5	I have found this process fairly labour intensive and resource for these tasks is generally not costed
No demand	2	no other groups have requested trial datasets to date
Risky task	5	it is a source of worry that we might inadvertently do something illegal
+ Positive	29[b]	
Committees/expert approvals	5	this process simply requires a robust documented process for deciding to release the data
Use of controlled environment	3	data should be released via a secure research environment
Use of data repositories	4	Use of institutional repository
Easy data transfer	6	Once an appropriate transfer method was identified and agreed, releasing the data set was straightforward
Easy process	7	Very easy. But if you trial is not controversial or current
Open access	1	This is not a concern if you are releasing open-access
Vetoing researchers	3	I feel comfortable releasing to bona fide researchers with good research ideas

[a]Only main nodes added (+)
[b]Only children nodes added

Table 4. Opinions about the process of maintaining datasets.

Name of code	No. of comments	Description/representative quote
Experience of maintaining	24[a]	In your opinion, how was your experience in the process of maintaining released de-identified/anonymised clinical trials datasets? (e.g. what have worked? what have not worked? any concerns? any assurances?)
+ No experience	9	No experience of this
+ No necessary	3	We don't maintain released datasets
+ Negative/burden	11[b]	
Bespoke process	1	there was no standardised way of maintaining de-identified datasets… anyone tasked with preparing a data release would first have to assess previous work. This was a large waste of time
Issue with contracts	2	Main issue has been regulatory/contracts rather than data process
Dynamic process	3	Using a controlled process means that we have much more control over the data regarding version control and ensuring that the data remains accessible using current technology
Difficult to keep links	1	Need to have correct processes in place to ensure that additional data can be linked with the existing data released without compromising data integrity or de-identification

(continued)

Table 4. (continued)

Name of code	No. of comments	Description/representative quote
Difficult to keep process for access	1	Some trial datasets that are still potentially available from studies I have worked on are password protected and the staff who knew the passwords have left without remaining staff having requested these.
No resources	3	It is difficult to arrange a process for maintaining controlled access that is available long-term and affordable
+ Positive/constructive	1[b]	
Keep for free by organisation	1	Kept in university

[a]Only main nodes added (+)
[b]Only children nodes added

Table 5. Views about re-identification risks scores.

	Participants N = 38 n (%)
Awareness of re-identification risk scores for assisting in the release of de-identified/anonymised clinical trial datasets	
I have never heard of them	18 (47)
I have heard about them, but I am not so sure what they are	9 (24)
I have a general understanding but do not use them	11 (29)
I have a good understanding and use them sometimes	0
I have a strong understanding and use them frequently	0
What are the barrier for not using the re-identification risk scores[a]	
Lack of funding	5 (45)
Lack of relevant training	11 (100)
Lack of time	8 (73)
Other[b]	4 (36)

[a]Only applicable to the 11 participants who answered in the previous question 'I have a general understanding, but do not use them'. Participants were allowed to mark multiple barriers.
[b]Other =1 'I'm not convinced they are necessary', 1 'Possible lack of benefit if following existing good process', 1 'Specific examples based around clinical trial data', 1 'study context, harder to conceptualise this metric than something descriptive. I'm fully aware of the irony that a statistician is essentially saying use free text here! I think they do have something to offer but need context'.
Note: As No one selected 'I have a good understanding, and use them sometimes' or 'I have a strong understanding, and use them frequently' in Q15, Q16–Q20 were no longer feasible and the survey automatically skipped to Q21 for the participant who answered 'I have a general understanding, but do not use them', and to Q22 for the participants who answered 'I have never heard of them' or 'I have heard about them, but I am not so sure what they are'

about them, but were not so sure what they are' and 'have a general understanding, but did not use them'. This last group expressed a lack of relevant training ((11/11) 100%), a lack of time ((8/11) 73%) and a lack of funding ((4/11), 45%) as the main barriers for not using re-identification risk scores (see Table 5). Further questions exploring the views about re-identification risks scores were not answered by any participant because they were only applicable if the participants answered that they had a 'good or strong' understanding of re-identification risk scores and used them 'sometimes or frequently'.

Regarding the wider aspects of re-identification risk, we attempted to identify which concerns related to the de-identified/anonymised datasets' properties were known to the researchers before release: 98% of the researchers always or often considered the data format, 71% always or often thought about the data uniqueness and 95% always or often contemplated the sensitivity of the data (Supplementary Figure 1). We also asked about the concerns around the release environment for de-identified/anonymised datasets and 52% of researchers always or often considered motivations to launch a re-identification attack on the datasets, 36% always or often considered the existence of auxiliary information to enable a re-identification attack, 26% always or often thought about the geographical location of the release, 61% contemplated consequences to individuals and 61% considered consequences to organisations if a successful re-identification attack occurred (Supplementary Figure 2).

Finally, we asked about any other aspects researchers considered before the release of anonymised/de-identified datasets. In total, 10 participants expressed concerns around the following themes: Any benefits of data release (3), contracts concerns (2), avoidance of open

Table 6. Any other aspects you consider before release.

Name of code	N	Description/representative quote
Any other aspects you consider before release	40[a]	Any other aspects you consider before you release de-identified /anonymised clinical trials data.
+ Concerns	10[b]	
Avoidance of open access	1	We've avoided true open access data release
Any benefits of data release	3	What are the potential scientific and societal benefits from release?
Contract concerns	2	Is the contract appropriate and does it make clear the duties of the requestor?
Ethical approval concerns	1	Ethics worries me
Expertise of data releaser	1	release is down to the highest data owner which, in our case, is the main trial statistician (with oversight). I wouldn't leave this decision to techies
Patients' expectation	1	Patient expectations of how their data would be handled (was consent obtained, how was the consent worded, what do patient representatives suggest about sharing that data specifically?)
Reasons to release data	1	(is it to be included in further meta-analysis, for potential secondary analysis, or are just released for 'accountability'/ because journals asked us to). What was said in the consent form about data use/sharing?
+ Steps/checks	30[b]	
Presence of any data uniqueness	4	Any release of special category data especially if there is a uniqueness inherent in the data
Execution of final checks	3	Robust checking of dataset prior to release to try to mitigate any of the risks
Verification of a good research question	8	favouring controlled release to groups that have a plan of what to do with the data
Provision of storage guidance for the released data	3	We give guidance on how the data should be stored when they are received, to reduce the risk of them being held insecurely
Consideration of the probability of re-identification	5	There is always a fear that you feel that the data, even with rigorous anonymous checks, by several people, that someone will be able to be identified with the data
Vetoing researchers	7	We ask for details of who the researchers are, and what their question is. We only release to bona fide researchers, with good research questions

[a]Only main nodes added (+)
[b]Only children nodes added

access (1), ethical approval concerns (1), expertise of data releaser (1), patients' expectation (1) and reasons to release data (1), and 30 participants mentioned steps/checks before data release which included verification of a good research question (8), vetoing researchers (7), consideration of the probability of re-identification (5), presence of any data uniqueness (4), execution of final checks (3) and provision of storage guidance for the released data (3) (Table 6).

Discussion

Researchers participating in this study belong to a group experienced in the conduct of clinical trials and in the processes associated with the preparation and release of de-identified/anonymised clinical trial datasets.

The dominant activities that respondents engaged in were the creation of the de-identified/anonymised datasets followed by the approval for release, while maintenance and peer review of the de-identified/anonymised datasets do not seem to be as common. This could be explained by the participant's profile, as statisticians and trial managers are more likely to take part in the creation of de-identified/anonymised datasets,[37,38] while IT specialists tend to be involved in the uploading and maintenance of the datasets, and this group is not represented in our results (despite the survey being open to it). Another possible explanation is that we assumed sharable datasets would be created at the end of a study, ready to be sent out on request.[39–41] However, in reality, sharable datasets are often created on demand, eliminating the need for maintenance.[42]

It is encouraging that 92% of the participants used some sort of documentation (either external or internal) for the de-identification/anonymisation process, with the process of how to de-identify/anonymise and release datasets being highly represented; conversely, the

implementation of risk evaluation is only modestly represented in these documents.

The most common overall process for the release of de-identified/anonymised clinical trial datasets, as reported by participants, was 'de-identification followed by anonymisation under controlled access', here clinical trial datasets are de-identified (key items stripped from the dataset), this is followed by data manipulation techniques to further anonymise the datasets and finally, datasets are released via the implementation of, for example, data sharing agreements, the location of the datasets behind secure access barriers or the identification and vetoing of secondary researchers and their research ideas. This process matches what we found in our previously published systematic scoping review.[43] This is promising because researchers are following the proposed recommendations/guidelines, which over time are providing a robust process, as evidenced by the fact that we do not yet have any known cases of a successful re-identification attack in the United Kingdom in clinical trial datasets.

When we explored researchers' opinions on the process of de-identifying/anonymising and maintaining datasets, negative sentiments seemed to dominate. Opinions on the data release process were balanced between positive and negative views. This suggests that de-identifying/anonymising the data is more troublesome than releasing it. This could be explained by the time in which data preparation and sharing activities are occurring. These activities tend to happen at the end of the studies, when the budget has been expended and the teams have pressures from other live projects.[37] The International Committee of Medical Journal Editors (ICMJE) have acknowledged this situation and they are recommending any trials that started enrolling participants after the 1 January 2019 must have a data sharing plan in the trial's registration.[41] However, at present, it might be premature to expect to see the emerging impact of that recommendation. Time is not the only constraint; it is known that de-identification and anonymisation of datasets are potentially complex and getting it wrong could have profound consequences.[44,45]

Regarding re-identification risk scores, it was expected that researchers would not know about how to calculate them, due to two main reasons: first, they are briefly described in the current guidance documents for clinical trials,[43] and second, the proposed re-identification risk scores come from health records management, so they are not common knowledge among clinical trialists. The small group of participants, who were aware of re-identification risk scores but were not using them, cited the primary reasons as a lack of training and time. Addressing the lack of training is an aspect that could be considered. This is an emerging topic within clinical trials where there are research and training gaps because researchers need a clear and tailored tool that they can use to estimate the re-identification risk for datasets.

We attempted to explore concerns about known parameters that could affect the re-identification risk,[46–50] using variables related to the datasets and to their release environments. We observed that even with no formal training, some researchers are already intuitively addressing these parameters and thinking of ways of mitigating their impact during the preparation and before the release of de-identified/anonymised datasets. Of course, robust guidance and availability of training could help to increase the level of engagement with the features that could affect re-identification risk.

Finally, we invited comment on issues that were not addressed in the rest of the survey, but notably the responses did not identify any new practices or concerns.

Comparison with existing literature

The report by Humphreys et al.[10] dealt with issues regarding wider aspects of data sharing for clinical trials and it commented that better guidance, more resources and training are required to fulfil data sharing from clinical trials. Our survey results agree with these findings. So, this is not an exclusive issue for de-identified/anonymised datasets.

Naudet et al.[51] reported a low incidence of sharing of de-identified/anonymised clinical trial datasets. They explained that the main reasons for the ICMJE data sharing policy not being implemented are lack of resources and training, lack of unified concepts (e.g. multiple definitions for anonymisation), real or perceived risk and the need to protect the interests of researchers and patients. This suggests that the barriers are not only researchers' opinions but also a reality. Currently, promises to eventually share data are not being kept.[42,52] Therefore, a future where data sharing is the norm is still out of reach until these issues are mitigated.

Humphreys et al.[10] highlighted re-identification risk as a key issue, but they did not describe how this risk should be calculated or quantify an acceptable level of risk. Our study is a first step to encourage a research stream for the underdeveloped area of re-identification risk estimation on clinical trial datasets.

Strengths and limitations

We sent the survey to mailing lists involving large numbers of people to maximise the chance of responses from individuals eligible for the survey. However, many of the people in these mailing lists would not have been eligible. We therefore could not investigate the response rate or the response bias of the survey. We collected responses from 38 participants, which, according to the

methodology used in this research, is sufficient to reach data saturation with respect to opinions, as we exceeded the minimum of between 12 and 30 survey responses. However, we may not have heard all opinions because, for example, we did not receive surveys from every UKCRC-registered CTU (population size n = 52). In addition, the CTUs represented in our sample may be a biased subset. We also do not know if the participants filling out the survey were speaking solely about their personal experience or if they were representing their CTU. However, the researchers who participated in this study were experienced and appeared to have relevant hands-on experience of the process of de-identification/anonymisation of clinical trial datasets. This gives strength to the results. However, eligible individuals self-identified as experienced in the subject matter and, we did not assess the details of this experience; instead, we inquired about the number of years of experience in de-identification/anonymisation.

All eligible individuals reached the end of the survey, and this is evidence of highly motivated participants interested in the topic of the survey. Such highly motivated individuals could potentially provide mostly positive experiences, and we might not have fully engaged with researchers who have done de-identification/anonymisation and have had adverse experiences or difficulties in this area. Nevertheless, we recorded a high incidence of negative sentiments.

As our resources were limited, we based the survey in the United Kingdom due to the complexities of ethical approvals. We could not predict where our potential international participants were going to be based and this restricted the ethical approval application to only the United Kingdom. Therefore, it was not possible to explore what is happening in other countries. To our knowledge, this has not been studied in other settings and future research addressing this gap would be valuable to confirm the generalisability of these findings. In this regard, we are sharing our full protocol and survey. Of course, many of the issues highlighted in this study are common global problems, so this UK study could be relevant to the wider research community.

The themes were manually coded and, therefore subjective; however, other reviewers sense-checked the coding and any disagreements were resolved via discussion within the research team.

This study is covering an emerging part of clinical trials research, for which even consensus about the definition of anonymisation does not exist.[43] To avoid misinterpretations in the survey, we provided definitions in the body of the survey to counter this issue. However, it cannot be ruled out that some of reported practices could be related to the sharing of pseudonymised data, as indicated by the use of controlled access.

Conclusion

It is positive to see that the majority of responders reported using documented processes for de-identification and anonymisation of clinical trial datasets. However, our survey results clearly indicate that there are still gaps in the areas of guidance, resources and training to fulfil sharing requests of de-identified/anonymised clinical trial datasets. In addition, the investigation of applications of re-identification risk scores on de-identified/anonymised clinical trial datasets could help with the development of an objective process to assess the re-identification risks and probabilities of re-identification attacks, which in turn can harmonise efforts towards more secure de-identified/anonymised datasets.

Meanwhile funders and sponsors should continue to foster and support activities regarding the preparation of de-identified/anonymised clinical trial datasets with the intention to share, such as training and funded time for these tasks.

Acknowledgements

The authors thank all the survey participants. They used the STROBE cross-sectional checklist when writing the report.[53] They thank the anonymous peer reviewers and the editors whose comments greatly contributed to the improvement of the article.

Authors' contributions

A.R., S.C.L., T.J. and C.J.W. conceived the idea for this work supported by S.E. A.R. wrote the first draft. All authors contributed to the protocol and to this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: A.R. has a scholarship from the University of Edinburgh to undertake a PhD with support from the Asthma UK Centre for Applied Research (AUKCAR grant no. AUK-AC-2012-01). S.C.L. and C.J.W. are supported in this work by their employment at the Edinburgh Clinical Trials Unit. T.J. is supported by Asthma UK as part of the Asthma UK Centre for Applied Research (grant nos. AUK-AC-2012-01 and AUK-AC-2018-01), S.E. is supported in this work by her employment at the Pragmatic Clinical Trials Unit. All of the authors contributed to protocol or article development. Neither sponsor (AUKCAR) nor funder (University of Edinburgh) contributed to protocol or article

Ethics approval

This project did not collect identifiable or personal participant data or personal sensitive information; therefore, this was a low-risk project, and we followed the ethical review processes coordinated by the Edinburgh Medical School Research Ethics Committee (EMREC). Protocol, survey and invitation letter submitted to the EMREC for consideration are in Additional Files 1, 2 and 4. We received a favourable ethical opinion (reference: 22-EMREC-027) on 1 June 2022 (Additional File 7).

ORCID iDs

Aryelly Rodriguez https://orcid.org/0000-0002-1352-3922
Tracy Jackson https://orcid.org/0000-0002-6188-3607
Christopher J Weir https://orcid.org/0000-0002-6494-4903

Availability of data and materials

All data collected for this study are included in this article as supplementary information files with the exclusion of the free-text data, which may be requested from the corresponding author for further reasonable research.

Supplemental material

Supplemental material for this article is available online.

References

1. Dal-Ré R. Access to anonymized individual participant clinical trials data: a radical change of mind by the most prestigious medical journals. *Arch Bronconeumol* 2018; 54(2): 65–67.
2. Pisani E, Aaby P, Breugelmans JG, et al. Beyond open data: realising the health benefits of sharing data. *BMJ* 2016; 355: i5295.
3. Bertagnolli M, Sartor O, Chabner B, et al. Advantages of a truly open-access data-sharing model. *N Engl J Med* 2017; 12: 1178–1181.
4. Clinical Study Data Request (CSDR). Clinical Study Data Request, 2020, https://clinicalstudydatarequest.com/
5. The Yale University. Yale University Open Data Access (YODA) Project, 2020, http://yoda.yale.edu/
6. Vivli Center for Global Clinical Research Data. Vivli, a global data-sharing and analytics platform, 2020, https://vivli.org/
7. El Emam K and Arbuckle L. *Anonymizing Health Data: Case Studies and Methods to get you Started*. Sebastopol, CA: O'Reilly Media, Inc., 2013.
8. Rodriguez A, Lewis S, Eldridge S, et al. *What are the re-identification risk scores of publicly available anonymised clinical trial datasets?*The University of Edinburgh, 20202, https://vivli.org/what-are-the-re-identification-risk-scores-of-publicly-available-anonymised-clinical-trial-datasets/
9. El Emam K. *Guide to the De-Identification of Personal Health Information*. Boca Raton, FL: CRC Press, 2013.
10. Humphreys GS, Merriott G, Knowles R, et al. *Clinical trial data sharing: What we've heard from researchers*. Figshare Report, 2020, https://wellcome.org/sites/default/files/clinical-trial-data-sharing-what-weve-heard-from-researchers.pdf
11. Creswell JW and Creswell JD. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th ed. New York: Sage, 2018.
12. Fink A. *The Survey Handbook*. New York: Sage, 2003.
13. Boynton PM and Greenhalgh T. Selecting, designing, and developing your questionnaire. *BMJ* 2004; 328: 1312–1315.
14. Stehr-Green PA, Stehr-Green JK, Nelson A, et al. Developing a questionnaire. *FOCUS Field Epidemiol* 2003; 2: 1–6.
15. U.S. Department of Health & Human Services (HHS). HHS – Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Washington, DC: *US Department of Health and Human Services*, 2012, https://wwwhhsgov/hipaa/for-professionals/privacy/special-topics/de-identification/indexhtml
16. Hrynaszkiewicz I, Norton ML, Vickers AJ, et al. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials* 2010; 340: c181.
17. Sweeney L. k-Anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems* 2002; 10: 557–570.
18. Dudovskiy J. *The Ultimate Guide to Writing a Dissertation in Business Studies: A Step-by-Step Assistance*. Pittsburgh, PA: Scientific Research, 2016, p. 51.
19. Lavrakas PJ. *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage, 2008.
20. Microsoft. MS Forms. 2016, p. *Part of Office 365*, https://www.microsoft.com/en-us/microsoft-365/online-surveys-polls-quizzes
21. UK Clinical Research Collaboration (UKCRC). Clinical Trials Units, https://www.ukcrc.org/research-infrastructure/clinical-trials-units/ (2023, accessed 05 Dec 2023)
22. UK Clinical Research Collaboration (UKCRC). Clinical Trial Units (CTUs) registered in the UK Clinical Research Collaboration (UKCRC) network, https://ukcrc-ctu.org.uk/registered-ctus/ (2020, accessed 30 October 2020).
23. PHUSE Limited. The Global Healthcare Data Science Community, https://phuse.global/ (2020, accessed 26 October 2020).
24. Allstat. Statistics email discussion list for the UK Education and Research communities, https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=allstat (2020, accessed 26 October 2020).
25. Hennink M and Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. *Soc Sci Med* 2022; 292: 114523.
26. Microsoft. Security and privacy in Microsoft forms, https://support.microsoft.com/en-us/office/security-and-privacy-in-microsoft-forms-7e57f9ba-4aeb-4b1b-9e21-b75318532cd9 (2020, accessed 30 October 2020).

27. Microsoft. Data storage for Microsoft Forms, https://support.microsoft.com/en-us/office/data-storage-for-microsoft-forms-97a34e2e-98e1-4dc2-b6b4-7a8444cb1dc3 (2020, accessed 30 October 2020).
28. The University of Edinburgh. Data – Data Services, https://www.ed.ac.uk/information-services/research-support/research-computing/ecdf/data (2020, accessed 30 October 2020).
29. The University of Edinburgh. Use University services, https://www.ed.ac.uk/infosec/information-protection-policies/procedures-guidance/use-university-services (2020, accessed 30 October 2020).
30. The University of Edinburgh. Working with sensitive data, https://www.ed.ac.uk/information-services/research-support/research-data-service/during/sensitive-data (2020, accessed 30 October 2020)
31. The University of Edinburgh. Data Protection Handbook, https://www.ed.ac.uk/sites/default/files/atoms/files/dataprotectionhandbookv10.pdf (2020, accessed 30 October 2020).
32. The University of Edinburgh. Policy and handbook, https://www.ed.ac.uk/data-protection/data-protection-policy (2023, accessed 05 November 2023).
33. SAS Institute Inc. SAS 9.4 [Computer software] TS level 1M4, Copyright © 2016 SAS Institute Inc. Cary, NC, 2013.
34. Braun V and Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006; 3: 77–101.
35. Gibbs GR. Thematic coding and categorizing. *Analyz Qual Data* 2007; 703: 38–56.
36. QSR International. NVIVO. Release 1.6.1 ed., 2022, https://help-nv.qsrinternational.com/20/mac/Content/about-nvivo/whats-new.htm
37. Tudur Smith C, Nevitt S, Appelbe D, et al. Resource implications of preparing individual participant data from a clinical trial to share with external researchers. *Trials* 2017; 18: 319.
38. Keerie C, Tuck C, Milne G, et al. Data sharing in clinical trials – practical guidance on anonymising trial datasets. *Trials* 2018; 19: 25.
39. Chan A-W, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 Statement: defining standard protocol items for clinical trials. *Revista Panamericana De Salud Pública* 2015; 38: 506–514.
40. Ohmann C, Banzi R, Canham S, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open* 2017; 7: eb018647.
41. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials. *BMJ* 2017; 357: j2372.
42. Kochhar S, Knoppers B, Gamble C, et al. Clinical trial data sharing: here's the challenge. *BMJ Open* 2019; 9: e032334.
43. Rodriguez A, Tuck C, Dozier MF, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: a scoping review. *Clin Trials* 2022; 19(4): 452–463.
44. The National Archives. Data Protection Act, 2018, https://www.nationalarchives.gov.uk/information-management/legislation/data-protection/#:~:text=Data%20protection%20law%20changed%20from,on%20the%20Information%20Commissioner's%20website.
45. The Crown Prosecution Service. Data protection act 2018 – Criminal Offences, 2018, https://www.cps.gov.uk/legal-guidance/data-protection-act-2018-criminal-offences
46. Dankar FK, El Emam K, Neisa A, et al. Estimating the re-identification risk of clinical data sets. *BMC Med Inform Decis Mak* 2012; 12: 66.
47. Xia W, Liu Y, Wan Z, et al. Enabling realistic health data re-identification risk assessment through adversarial modeling. *J Am Med Inform Assoc* 2021; 28: 744–752.
48. Simon GE, Shortreed SM, Coley RY, et al. Assessing and minimizing re-identification risk in research data derived from health care records. *Egems* 2019; 7: 6.
49. Jiang Y, Mosquera L, Jiang B, et al. Measuring re-identification risk using a synthetic estimator to enable data sharing. *PLoS ONE* 2022; 17(6): e0269097.
50. Taneja H and Singh AK. Preserving privacy of patients based on re-identification risk. *Procedia Computer Science* 2015; 70: 448–454.
51. Naudet F, Siebert M, Pellen C, et al. Medical journal requirements for clinical trial data sharing: ripe for improvement. *PLoS MED* 2021; 18(10): e1003844.
52. Strom BL, Buyse ME, Hughes J, et al. Data sharing – is the juice worth the squeeze? *N Engl J Med* 2016; 375: 1608–1609.
53. Von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007; 370: 1453–1457.

5.3 Conclusion

This part of the thesis explored the views of UK clinical trials researchers regarding their experiences with creating and releasing anonymised datasets, generating and using re-identification risk scores, and their perspectives on broader re-identification risks. Re-identification risk scores, as described by El Emam (El Emam 2013) , use information in anonymised datasets to estimate the probability of identifying individuals. While these scores are currently used primarily for health records, they could also be applied to clinical trials, as discussed in Chapter 4. Although there may be instances where re-identification risk scores have been used in clinical trials, such applications are not yet widely documented.

The research in Chapter 5, aimed to understand what is working and what is not in the creation and release of anonymised datasets and to determine if re-identification risk scores are used by UK clinical trials researchers. To achieve this, I employed an online exploratory cross-sectional descriptive survey, including both open-ended and closed questions, to gather detailed insights. The survey, conducted in English, was structured in five parts: consent and eligibility check, researchers' work background, experiences with anonymised datasets, knowledge and use of re-identification risk scores, and views on re-identification risks.

Participants were recruited from various UK-based clinical trial units, data transparency groups, and an academic conference (6th ICTMC 2022 in Harrogate UK) (International Clinical Trials Methodology Conference (ICTCM), Sydes et al. 2023) , using convenience non-probability sampling. The survey was active from June to October 2022, resulting in 52 responses to invitation, of which 38 were eligible to participate. Most participants were statisticians or senior statisticians, with significant experience in their roles.

The survey results showed that the most common involvement with anonymised

datasets was in their creation and approval for release. Documents and guidance were used to inform the process, and the most used method for releasing datasets involved de-identification followed by anonymisation, with their release under controlled access, this mirrors the most recommended method as per our scoping review(Rodriguez, Tuck et al. 2022) .

Thematic analysis revealed that most opinions on the process of de-identifying and anonymising datasets were negative, citing long processes, lack of advice, and data constraints. Positive sentiments included straightforward processes and sharing with trusted researchers. Regarding the release of datasets, opinions were more balanced between positive and negative views.

The survey also found that none of the participants used re-identification risk scores, with many unaware of what they are or lacking the training to use them. Researchers did consider various factors before releasing datasets, such as data format, uniqueness, sensitivity, and potential re-identification risks. However, there is a need for more robust guidance and training in this area. At the time of writing this thesis (August 2024), the article associated with this chapter has been viewed and downloaded 250+ times (Clinical Trials, a SAGE journal, 2022 (Clinical Trials a SAGE journal 2022)).

This study highlights the need for improvements in guidance, resources, and training for the de-identification and anonymisation of clinical trial datasets. While researchers follow recommended processes, there are still significant challenges and gaps in the implementation and understanding of re-identification risks. Future research should address these gaps to improve data sharing practices and ensure the safe and effective use of anonymised clinical trial datasets. In the next and final chapter, I will amalgamate the findings from Chapter 3, 4 and 5 to produce an overarching conclusion for my PhD research.

Chapter 6 Overall Discussion and Conclusions

6.1 Introduction

This thesis discusses the increasing emphasis on sharing anonymised clinical trial datasets, driven by publishers and funders. This practice is essential for disseminating current research, enabling new investigations, and maximising scientific efforts. However, anonymising data is complex, and achieving robust anonymisation can lead to a loss of detail necessary for comprehensive analysis. Therefore, there is a balance between minimising the risk of re-identification and maintaining sufficient detail to replicate original analyses and answer further, valid, research questions.

This final chapter relates to the fifth objective of the PhD: “To develop evidence-based recommendations on anonymisation techniques and data security for clinical trial datasets”. To this effect, I first summarise the findings in Chapters 3, 4 and 5. I then discuss the implications of these findings for current practices and their connections with recent developments in the field, and outline directions for future research. Finally, I provide the overarching conclusion of this 6-year part-time PhD.

6.2 Summary of findings

There are growing demands for sharing anonymised datasets from clinical trials within the scientific community, amidst varying recommendations on anonymisation. To begin this PhD research, I conducted a systematic scoping review on MEDLINE®, EMBASE, and Web of Science™, including any publication with recommendations on anonymisation for data sharing from clinical trials, to understand what was available to researchers that was tailored to clinical trials datasets. Fifty-nine articles were included, which identified three main concepts: anonymisation, de-identification, and

pseudonymisation. Commonly suggested anonymisation techniques involved removing direct patient identifiers and modifying indirect identifiers to minimise identification risks. This first part of the research found no single standardised anonymisation recommendation but noted a trend towards consensus on definitions and techniques and highlighted the necessity of pairing anonymisation with controlled access to protect patient privacy.

The second part of this PhD investigated whether publicly available anonymised clinical trial datasets pose a privacy risk to participants. I accessed a broad sample of de-identified/anonymised datasets from human participants and classified personal data within these datasets. I calculated re-identification risk scores based on combinations of indirect identifiers such as sex, age, and race. These scores ranged from 0 (indicating the lowest risk) to 1 (maximum risk). Seventy datasets were analysed, with 31 shared openly and 39 under controlled access. The datasets had an average of four identifiers, with average risk scores ranging from 0.47 to 0.91. I found out that clinical trial datasets contained rich personal details, highlighting the importance of carefully managing re-identification risks. I suggest using re-identification risk scores to guide the anonymisation process for secondary research purposes. This recommendation is incorporated into a comprehensive strategy, which is detailed in the flow chart presented in Figure 1, Section 4.2 of this thesis.

The third and final part of this PhD explored the views and experiences of UK clinical trial researchers regarding de-identification, anonymisation, data release methods, and re-identification risk estimation for clinical trial datasets. An online survey of UK clinical trials researchers was conducted between June 2022 and October 2022, yielding 38 responses. The majority (35, 92%) used internal documentation or published guidance for de-identification/anonymisation, with controlled access being

the most common release method. Despite some familiarity with re-identification risk estimation, it was rarely used. Surveyed researchers reported negative experiences due to a lack of resources, guidance, and training. This part of the PhD concluded that while documented processes are commonly used, there are significant gaps in guidance, resources, and training for sharing de-identified/anonymised datasets, and re-identification risk estimation remains underdeveloped.

6.3 Context and implications

The landscape of clinical trial data sharing is rapidly evolving, making it increasingly difficult to justify the refusal to share such data. Several recent developments underscore this shift. This section explores these developments, the implications of my research, and how they intersect.

In 2019, the University of Bristol was ordered by a UK First-tier tribunal to release anonymised clinical trial data collected on 100 children in the Bath area (Healy 2019). Additionally, funders like the MRC are already indicating the imperative nature of data sharing at the funding stage with a robust data sharing policy (Medical Research Council (MRC) 2023). A significant recent development in this area is the new policy introduced by the BMJ, effective from the 5th of March 2024. The BMJ now mandates that all authors of submitted research trials must share relevant trial data in a publicly accessible, enduring repository, such as Vivli (Vivli Center for Global Clinical Research Data 2020), before publication. Moreover, authors are required to submit the analytical code used in their research as a supplementary file, which will be permanently accessible alongside the published paper (Loder, Macdonald et al. 2024). This policy aims to enhance transparency and scrutiny of medical research by ensuring that data and code are readily available for re-analysis and verification by other researchers. The BMJ might

be taking this step to address concerns observed by Butte (Butte 2021), Gabelica et al. (Gabelica, Bojčić et al. 2022) and Esmail et al. (Esmail, Kapp et al. 2023) that even when researchers committed to data sharing via a "data sharing statement," this did not ensure the data was actually accessible, and more needed to be done to increase data availability. In general terms, it could have been predicted that was going to be next step, the requirement of a mandatory, permanently accessible dataset in a defined location before publication, rather than relying on authors to self-regulate and fulfil their promises in their data sharing statements. It seems likely that other publishers and funders will follow suit in requiring mandatory data sharing. Moreover, reliable data is crucial for advancing AI research and its applications, which have the potential to benefit society significantly. For example, AI has already led to the discovery of a new antibiotic (Gallagher 2023). While making clinical trial data available may not yet be an explicit duty, it could significantly contribute to AI development and enhance the value of these datasets (Kulakiewicz, Parkin et al. 2022), aligning with emerging trends in research transparency and data sharing. However, it is important to approach this with careful consideration. We should ensure that data is shared responsibly and used intelligently, as there is potential for misunderstandings or misinterpretations of data fields. Establishing robust guidelines and oversight is crucial to manage how data is used in AI initiatives, ensuring that it supports meaningful and accurate advancements without compromising the integrity of the research.

The growing attention to data sharing is further reflected in the increasing importance placed on anonymisation. The principles discussed by Hrynaszkiewicz et al. (Hrynaszkiewicz, Norton et al. 2010) regarding anonymisation in clinical trials are gaining traction (Figure 6-1), much like how HIPAA regulations have become widespread in the US.

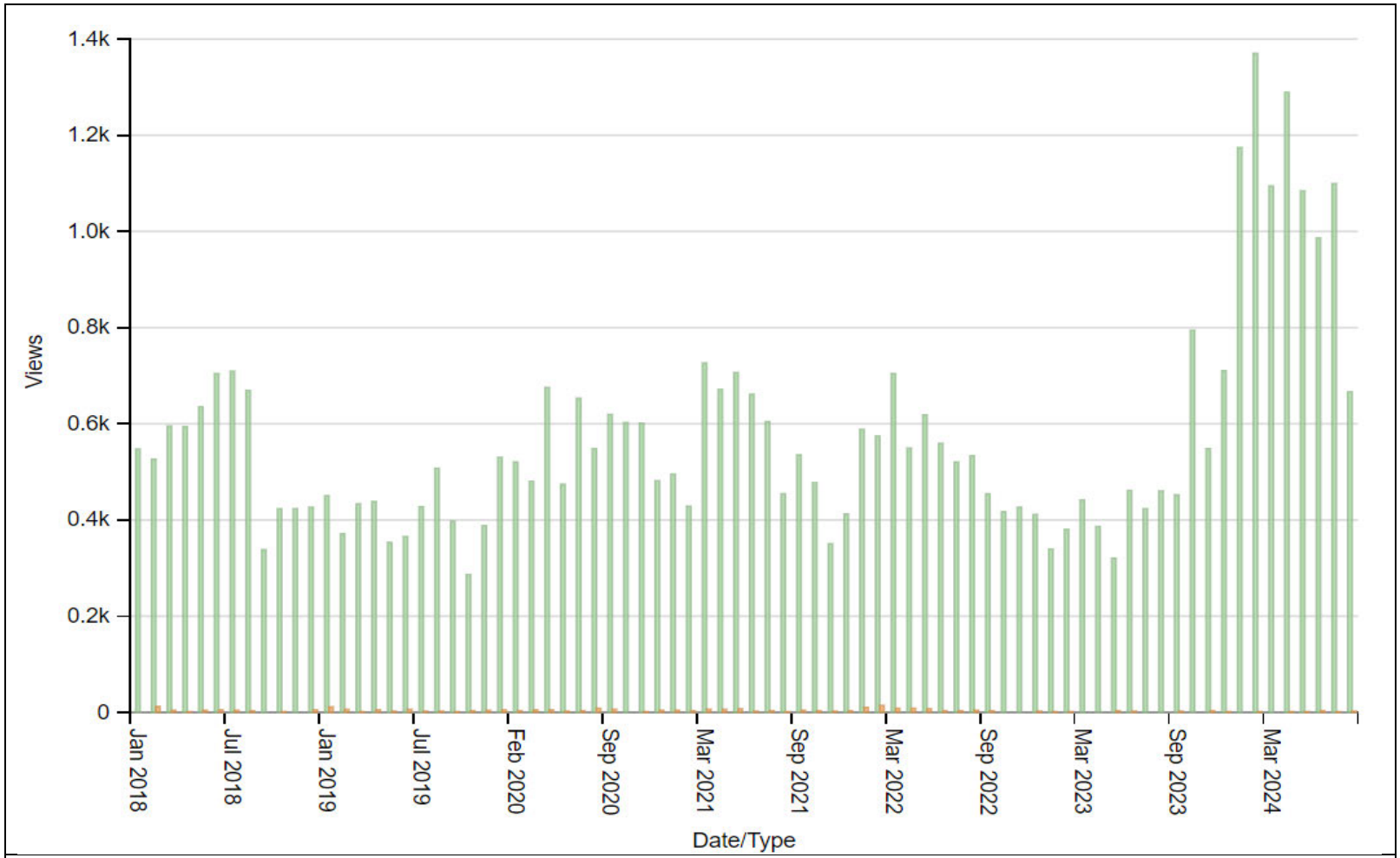


Figure 6-1 Usage statistics for Hrynaszkiewicz et al. 2010 – January 2018 – August 2024

Image source: https://www.bmj.com/content/340/bmj.c181/article-info

Bearing in mind the central role that sharing clinical trial data will play in years to come, and considering anonymisation one of the key solutions to improve data sharing(Hrynaszkiewicz, Simons et al. 2020), we, trialists, need to adopt unified working concepts. The findings from the first part of this PhD provide evidence that there are numerous definitions of anonymisation and de-identification. Many issues could be resolved if we agreed on universal definitions; however, we have to understand that this agreement must come from the community rather than from regulators(El Emam and Abdallah 2015) (Law, Couturier et al. 2022). The law has to cater for a wide range of datasets and has to implement a one-size-fits-all approach, but this may not be suitable for all fields. The good news is that Hrynaszkiewicz et al. (2010) (Hrynaszkiewicz, Norton et al. 2010) and Tudur Smith et al. 2015 (Tudur Smith, Hopkins et al. 2015) have already provided us with definitions, and it is time to implement them. For example, a new challenge has emerged, such as the confusion between personal data and anonymised data, particularly when asking participants to share their anonymised data. While it is correct to notify clinical trial participants that their data is going to be anonymised and shared for secondary purposes, it is contradictory to ask for their permission to share anonymised data, which is no longer personal data(El Emam, Hintze et al. 2020). This type of confusion stems from the lack of unified definitions. Although it might be unrealistic to call for a new single source of definitions, rallying around Hrynaszkiewicz et al. (2010) and Tudur Smith et al. (2015) could be a viable solution. Without moving towards accepted universal definitions, these kinds of challenges will continue to manifest.

Another relevant finding in this PhD is the confirmation that re-identification risks are never going to be zero, regardless of the steps taken to anonymise a dataset (Stalla-Bourdillon and Knight 2016) (El Emam 2013). Therefore, it is essential to educate ourselves

and the wider community with these key messages: "fully" anonymised data does not exist, rather, anonymisation exists on a spectrum (Figure 6-2), and it is a balancing act between data utility and re-identification risk minimisation, as confirmed by this research.

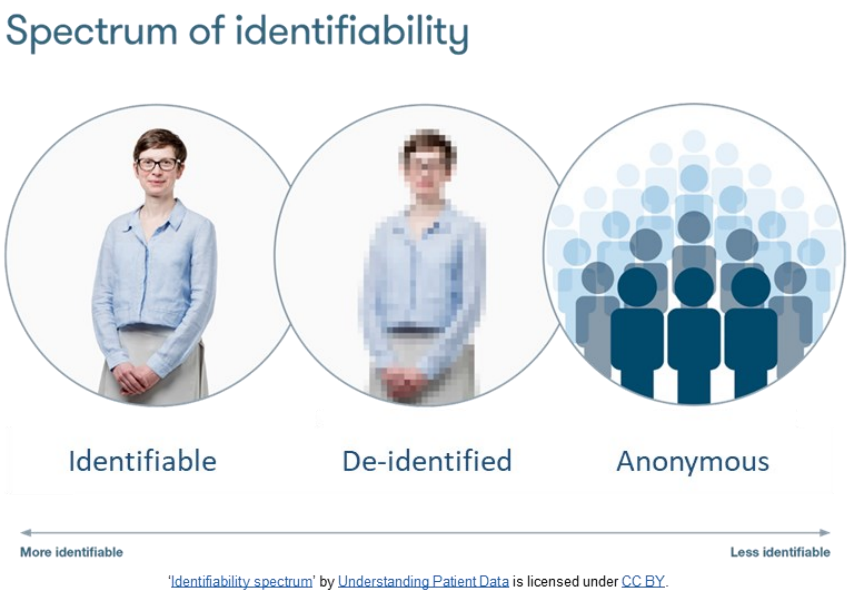


Figure 6-2 Spectrum of identifiability
Image source: https://understandingpatientdata.org.uk/weighing-up-risks#privacy
This image is openly licensed via CC BY 4.0. (https://creativecommons.org/licenses/by/4.0/)

We should consider adopting terms like “sufficiently anonymised” (El Emam, Rodgers et al. 2015) data, thus accepting some level of risk, having the confidence that in the UK, an attempt to re-identify such datasets is a criminal offence under the UK Data Protection Act 2018 (The Crown Prosecution Service 2018). Moreover, we should not fear those risks posed by anonymised data; instead, they should be studied and compared against the actual background risk we already endure due to our daily digital lives. For example, in 2018, the Cambridge Analytica scandal exposed how the Facebook users’ personal data was exploited (Trezza 2023); in 2020, personal information from 284 NHS Highlands patients was accidentally disclosed (BBC News 2020); in 2023, almost half a million people had their personal data stolen from the

Universities Superannuation Scheme (USS) via a cyber-attack (Davies 2023); and also in 2023, an NHS secretary illegally accessed personal medical records out of curiosity (Khan 2023). These examples illustrate that the greatest risks to our personal data perhaps come from the internet, the people who handle it, and the level of investment and protection afforded to it.

These constantly changing everyday exposures pose greater risks to privacy, offering attackers simpler and cheaper methods of obtaining personal data than acquiring anonymised datasets and attempting to re-identify them. Motivated attackers have easier targets to exploit before investing lots of considerable resources to explore anonymised datasets (Yakowitz 2011), as I did in the second part of this PhD. Successful re-identification attacks on anonymised dataset have occurred on data that have not been properly anonymised (El Emam, Jonker et al. 2011). Therefore, the risk is real if anonymisation is not done robustly. However, when anonymisation is properly executed and thoroughly checked, the chances of successful re-identification attacks could be small (IOM (Institute of Medicine) 2015).

Like data sharing, re-identification risk assessment is proving to be an ever-moving target. For example, Appendix 8 - Case Study - Professor Sweeney's Research highlights how Professor Sweeney's 20 years of work on re-identification risk has changed and evolved over time. Re-identification risk assessment is already recommended as good practice by high-level guidance such as the "Anonymisation: managing data protection risk code of practice" issued by the UK Information Commissioner's Office (Information Commissioner's Office (ICO) 2012) which aims to "detect and deal with re-identification vulnerabilities". The Institute of Medicine (IOM) (IOM (Institute of Medicine) 2015) also recommend its use, although their recommendation is more abstract. Specifically, the IOM describes re-identification

risk assessment at a “conceptual level”, meaning it outlines the importance and general principles of the assessment without delving into detailed, practical implementation methods. Their focus is on the theoretical framework and broader rationale for its use. The second part of this PhD provides evidence that assessing the re-identification risk in anonymised clinical trial datasets is not only feasible but can also inform the anonymisation process, offering a valuable entry point for the clinical trials community. This approach enhances transparency and builds confidence in the effectiveness of anonymisation practices. Existing documentation only shows theoretical examples of re-identification risk calculations, while this PhD takes a step further by providing calculations from real clinical trial datasets, offering valuable guidance and comparison for the clinical trial community.

Finally, it is essential to listen to the voices in the community and address the negative perception surrounding the sharing of clinical trial datasets. The third part of this PhD showed that more training and resources are required to adequately anonymise datasets and explore their re-identification risks, because currently the rewards of data sharing are few, but the efforts are substantial, as observed by Humphreys et al. (Humphreys, Merriott et al. 2020) Rios et al. (Rios, Zheng et al. 2020) and Modi et al. (Modi, Kichenadasse et al. 2023). However, the potential benefits to science and society make this an endeavour worth pursuing.

6.4 Future research

The topic of anonymisation of clinical trial datasets for secondary research is vast and controversial, and one PhD alone cannot do it justice. While data sharing of clinical trials is part of a broader conversation, progressing the issues around

anonymisation could help remove a significant barrier related to privacy protection (van Panhuis, Paul et al. 2014).

6.4.1 Enhancing current work

If I had more time, I would have liked to convert the scoping review in Chapter 3 into a living document that is continuously updated as new guidance becomes available. Additionally, a qualitative assessment of the studies included in the scoping review could have been valuable. Identifying a hierarchy of studies could facilitate progress towards unified concepts, eliminating issues of contradiction within the same concept.

6.4.2 Scoping reviews and anonymisation methods

The clinical trials community would greatly benefit from a comprehensive scoping review to identify and critically assess the most effective techniques for anonymisation and privacy models for clinical trial datasets; this was not covered by my scoping review. This review should also systematically search for re-identification assessment methods to ensure no advanced or superior methods beyond those by El Emam (2013) (El Emam 2013) are overlooked.

6.4.3 Re-identification risk scores and assessment

Regarding re-identification risk scores, it is worth exploring whether there should be maximum limits recommended for clinical trials datasets. Chapter 4 highlighted the dual aspects of re-identification risk assessment: a quantitative part (re-identification risk scores) and a qualitative part (prioritising risks based on probability and impact, considering motives for attacks and the data release context) (IOM (Institute of Medicine) 2015); (Stalla-Bourdillon and Knight 2016). I plan to conduct a small study, possibly via a cross-sectional survey, to collect perceived re-identification risks and

catalogue them with a group of experts using a rational approach to record risks and consider mitigating actions, like in Table 6-1.

What are the re-identification risks?	Who might be harmed and how?	What are we already doing to control the risks?	What further action do we need to take to control the risks?	Who needs to carry out the action?	When is the action needed by?	Done

Table 6-1 Proposed table for identification of qualitative re-identification risks

Source Adapted from Health and Safety Executive "Risk Assessment template" located at https://www.hse.gov.uk/simple-health-safety/risk/risk-assessment-template-and-examples.htm

I also aim to develop a training course and materials on calculating re-identification risk scores, as outlined in Chapter 4, to be delivered as a free webinar and recorded for future training resources. This initiative could start in the UK via UKCRC.

6.4.4 Assessing data utility

Future investigation should focus on assessing the utility of anonymised datasets. Chapter 4 included only a small number of datasets with anonymisation details, leaving this research without an evaluation of how different the anonymised datasets were from their non-anonymised versions. This comparison is crucial for assessing utility. Potential approaches include the "non-uniform entropy" recommended by El Emam and Arbuckle (El Emam and Arbuckle 2013) and a "generic method" investigated by Prasser et al. (Prasser, Bild et al. 2016), but their feasibility in clinical trials requires evaluation.

Calculating utility metrics necessitates access to non-anonymised datasets, meaning this assessment can only be conducted by the creators of the anonymised datasets. The exploration of utility will likely follow a path similar to re-identification risk

assessment, encompassing both quantitative and qualitative aspects. Ongoing efforts, such as Pilgram's research on the privacy-utility trade-off in anonymised data (Pilgram, Meurers et al. 2024), will continue to advance the field. However, the qualitative aspect remains to be explored.

6.4.5 Alternatives solutions to anonymisation.

While the concept of "sufficiently anonymised" data, as discussed in 6.3, presents a viable pathway for managing data privacy, it is not without its challenges, particularly in balancing the trade-off between data utility and privacy. As an alternative to traditional anonymisation techniques, the use of synthetic data has emerged as a promising approach to address these challenges (Azizi, Zheng et al. 2021, Jiang, Mosquera et al. 2022) .

Synthetic data is artificially generated data that mimics the statistical properties and distribution of real-world data but does not contain any actual individual-level information (Gretel Labs 2024). This method offers a way to preserve the privacy of individuals while still enabling analysis and research.

Among the advantages of synthetic data are its enhanced privacy, as the absence of real individual observations significantly reduces the risk of re-identification. This makes it an attractive option in scenarios where the anonymisation of real data is insufficient or where the legal implications of a privacy breach are severe. Furthermore, synthetic data can be shared more freely with external researchers or third parties, avoiding many of the privacy concerns associated with real data. It also offers an easier pathway to meeting regulatory requirements, as it does not link back to any real individuals, thereby simplifying the legal framework around data sharing and use.

However, it is important to note that one of the main drawbacks of synthetic data is that it may not capture all the nuances of the original data, particularly in complex datasets. The utility of synthetic data depends on how accurately it replicates the relationships and patterns present in the real data. Additionally, the generation of high-quality synthetic data requires advanced techniques and can be resource-intensive. The process also demands careful validation to ensure that the synthetic data retains the necessary properties for meaningful analysis. Finally, as synthetic data does not correspond to real events or individuals, findings based solely on synthetic data may lack real-world validation, potentially limiting their applicability.

Therefore, the choice between sufficiently anonymising real data and using synthetic data involves weighing the trade-offs between privacy, data utility, and practicality (Bamford, Lyons et al. 2022). While sufficiently anonymised data remains an important approach, especially given the legal protections against re-identification in jurisdictions like the UK, synthetic data offers a complementary or alternative pathway that may be particularly valuable in high-risk scenarios or where greater flexibility in data sharing is required. Future research should explore the conditions under which each approach is most appropriate, potentially leading to hybrid solutions that leverage the strengths of both methodologies.

6.4.6 Expanding survey participation

The survey in Chapter 5 could be expanded to include a larger and more diverse group of participants, including those from outside the UK, to improve the generalisability of the findings. Additionally, this survey could serve as a tool for assessing the impact of re-identification risk scores training courses.

6.4.7 Additional Resources investigation

Achieving appropriate anonymisation is complex, with significant barriers such as limited funding, resource allocation, and access to expert guidance. Tudur Smith et al. (2017) conducted important work in estimating the costs and time required to prepare a clinical trial dataset for sharing. However, as demonstrated in the process outlined in Chapter 4, there are many additional resources to consider. For example, institutions must address significant research governance elements when sharing data under controlled access models.

Given that Clinical Trials Units (CTUs) often lack the resources to perform extensive data manipulation or engage in prolonged legal consultations once a study has concluded, they may struggle with the practical aspects of preparing data for sharing or anonymisation due to limited post-study funding and personnel. It is crucial to acknowledge that CTUs frequently operate under tight financial and resource constraints, which hinders their ability to fully engage in the detailed processes required for effective data anonymisation and legal compliance.

CTUs need easy access to experts in research governance and legal matters who can provide guidance on whether data is "anonymised enough" for specific purposes. This support is vital because anonymisation is not merely a technical process; it also involves complex legal and ethical considerations that are often context-dependent. The current system may lack adequate support structures, making it difficult for clinical trial researchers to navigate the complexities of data anonymisation on their own. Further research into all the resources required would be beneficial.

6.4.8 Engaging patient and public Involvement (PPI) members

In 2021, I sought PPI members to help shape the qualitative component of the re-identification risk score. Although the results of my PhD focus on deliverables for trialists, the ultimate goal is the protection of clinical trial participants. The Asthma UK Centre for Applied Research (AUKCAR) PPI group has the expertise to help shape this aspect of the research. Figure 6-3 shows the recruitment advert seeking interested parties among the AUKCAR PPI group.

Figure 6-3 Advert for PPI

Image source: Original

Note: This was accompanied by an email inviting participation in a meeting about risk calculation, which was held on 27th Jan 2021

Five members of the AUKCAR PPI group accepted the invitation and met with me on 27th January 2021 via Zoom. After presenting the outline of the research detailed in Chapter 4 of this thesis, all five members expressed a desire to further engage with this study. They offered to provide advice on the findings related to re-identification risk scores and agreed to meet again once these results are available. The aim is to initially explore the implications of these findings for participants and to identify potential avenues for qualitative research that could be pursued. Due to delays in data analysis, I did not manage to meet with them before the end of my PhD, but I will do so as soon as the related article is published. This could complement findings by Howe (Howe, Giles et al. 2018) in her thesis "Participants' Attitudes Towards Data Sharing" (Howe 2021), where she explored clinical trial participants' views on data sharing. It would be interesting to approach participants with actual re-identification scores and gather their perspectives.

6.4.9 Addressing stigmatising conditions

Once anonymisation and re-identification risk concepts are unified and established for clinical trial datasets, research should evaluate these aspects for datasets studying stigmatising conditions, such as addictions, where data sharing practices are practically non-existent (Vassar, Jellison et al. 2020).

6.4.10 regulatory considerations

Finally, there should be more research to motivate regulators to explore stronger laws to penalise criminal activity and more favourable views for organisations that have taken all necessary steps to anonymise and share data robustly.

6.5 Overall Conclusion

This PhD research underscores the critical importance and complexity of sharing anonymised datasets from clinical trials. The findings reveal significant challenges, opportunities, and necessary actions to enhance data sharing practices while safeguarding participant privacy. The journey towards effective and safe sharing of anonymised clinical trial datasets is complex and requires careful consideration of privacy risks, data utility, and regulatory compliance. While fully anonymised data may never be achievable, "sufficiently anonymised" data represents a pragmatic goal.

To advance data sharing practices, the clinical trials community must embrace unified concepts and standards, supported by robust training and resources. Re-identification risk assessment should be integrated into the anonymisation process, as recommended by high-level guidance and demonstrated by this research. However, it is important to recognise that achieving appropriate anonymisation is not a straightforward process that can simply be implemented. In reality, significant barriers remain, particularly concerning funding, resource allocation, and the availability of expert guidance. These challenges underscore the need for more sustainable funding models and better access to expertise to ensure that anonymisation efforts are both effective and feasible.

This PhD provides a foundational understanding and actionable recommendations to enhance the anonymisation process and promote the responsible sharing of clinical trial data, ultimately benefiting scientific progress and public health.

Appendices

Appendix 1 - Chapter 3 - Associated "Self-Audit Checklist for Level 1 Ethical Review"

University of Edinburgh,
Usher Institute of Population Health Sciences and Informatics
RESEARCH ETHICS SUBGROUP

Self-Audit Checklist for Level 1 Ethical Review for PGR projects

See **Intra** website for further information: http://www.cphs.mvm.ed.ac.uk/intra/research/ethicalReview.php

NOTE to student: *Completion of this form should be under the* ***oversight*** *of your supervisor. A good strategy would be to complete a draft as best you can, then discuss with your supervisor before completing a final copy for your supervisor to sign.*

Proposed Project *(State research question and topic area, and briefly describe method/ data. Specify also countries in which data will be collected.):*

We aim to identify, describe and synthesise the methods/recommendations currently being used by researchers to anonymise datasets from clinical trials

There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. There are various sets of recommendations on how to perform anonymisation prior to sharing clinical trial data. We aim to systematically identify, describe and synthesise these recommendations. We will systematically search literature databases and websites of key organisations in the field. Any article reporting recommendations on anonymisation to enable data sharing in clinical trials will be included. There would not be a geographical restriction for identified sources. Two reviewers will independently screen titles/abstracts and full texts for eligibility. One reviewer will extract data from included papers and sense check by a second reviewer. Results will be summarised by narrative review. This scoping review will provide information about existing recommendations for anonymising clinical trial datasets in order to make them available for sharing and it will inform (if applicable) the development of new recommendations

1. Bringing the University into disrepute

Is there any aspect of the proposed research which might bring the University into disrepute? ~~YES~~/ NO

2. Data protection and consent

Are there any issues of DATA PROTECTION or CONSENT which are NOT adequately dealt with via established procedures? ~~YES~~/ NO

These include well-established sets of undertakings. For example, a 'No' answer is justified *only if*:

(a) There is compliance with the University of Edinburgh's Data Protection procedures (see www.recordsmanagement.ed.ac.uk);

(b) Respondents give consent regarding the collection, storage and, if appropriate, archiving and destruction of data;

(c) Identifying information (eg consent forms) is held separately from data;

(d) There is Caldicott Guardian approval for (or approval will be obtained prior to) obtaining/ analysing NHS patient-data.

(e) There are no other special issues arising about confidentiality/consent.

3. Study participants

a) Will a study researcher be in direct contact with participants to collect data, whether face-to-face, or by telephone, electronic means or post, or by observation? (eg interviews, focus groups, questionnaires, assessments) ~~YES~~/ NO

b) **Answer this only if qu. 3 above = 'YES':**
In ethical terms, could any participants in the research be considered to be ***'vulnerable'****?*
e.g. children & young people under age of 16, people who are in custody or care (incl. school), a marginalised/stigmatised group

Please tick one: 'vulnerable' ☐ not 'vulnerable' ☐

4. Moral issues and Researcher/Institutional Conflicts of Interest

Are there any SPECIAL MORAL ISSUES/CONFLICTS OF INTEREST? ~~YES~~/ NO

(a) An example of conflict of interest for a researcher would be a financial or non-financial benefit for him/herself or for a relative of friend.

(b) Particular moral issues or concerns could arise, for example where the purposes of research are concealed, where respondents are unable to provide informed consent, or where research findings could impinge negatively/ differentially upon the interests of participants.

(c) Where there is a dual relationship between researcher and participant (eg where research is undertaken by practitioners so that the participant might be unclear as to the distinction between 'care' and research)

5. Protection of research subject confidentiality

Are there any issues of CONFIDENTIALITY which are NOT adequately handled by normal tenets of confidentiality for academic research? ~~YES~~/ NO

These include well-established sets of undertakings that should be agreed with collaborating and participating individuals/organisations. For example, a 'No' answer is justified *only if*:

(a) There will be no attribution of individual responses;

(b) Individuals (and, where appropriate, organisations) are anonymised in stored data, publications and presentation;

(c) There has been specific agreement with respondents regarding feedback to collaborators and publication.

6. Potential physical or psychological harm, discomfort or stress

(a) Is there a FORSEEABLE POTENTIAL for PSYCHOLOGICAL HARM or STRESS for participants? ~~YES~~/ NO

(b) Is there a FORSEEABLE POTENTIAL for PHYSICAL HARM or DISCOMFORT for participants? ~~YES~~/ NO

(c) Is there a FORSEEABLE RISK to the researcher? ~~YES~~/ NO

Examples of issues/ topics that have the potential to cause psychological harm, discomfort or distress and should lead you to answer 'yes' to this question include, but are not limited to:

relationship breakdown; bullying; bereavement; mental health difficulties; trauma / PTSD; violence or sexual violence; physical, sexual or emotional abuse in either children or adults.

7. Duty to disseminate research findings

Are there issues which will prevent all relevant stakeholders having access to a clear, understandable and accurate summary of the research findings if they wish?* ~~YES~~/ NO

** If, and only if, you answered 'yes' to 3 above, 'stakeholders' includes the participants in the research*

Overall assessment

- If every answer above is a definite NO, the self-audit has been conducted and confirms the ***ABSENCE OF REASONABLY FORESEEABLE ETHICAL RISKS*** – *please tick box* ✓

  *This means that regarding this study, as currently self-audited, no further ethical review actions are required within Usher. However, if in the coming weeks/months there is any change to the research plan envisaged now (and outlined above), the study should be **re-audited** against a Level 1 form, because it may be that the change made negates the absence of ethical risks signed off here.*

  Two copies of this form should be taken for inclusion in the final dissertation/thesis and **the original should be returned to Usher Ethics admin** – usher.ethics@ed.ac.uk

  Receipt of this form will be acknowledged, but no formal letter of approval can be/will be sent. *If some formal letter is required for a funder or collaborator, please be sure to ask in your covering email, and we will send a form of words explaining the self-audit process.*

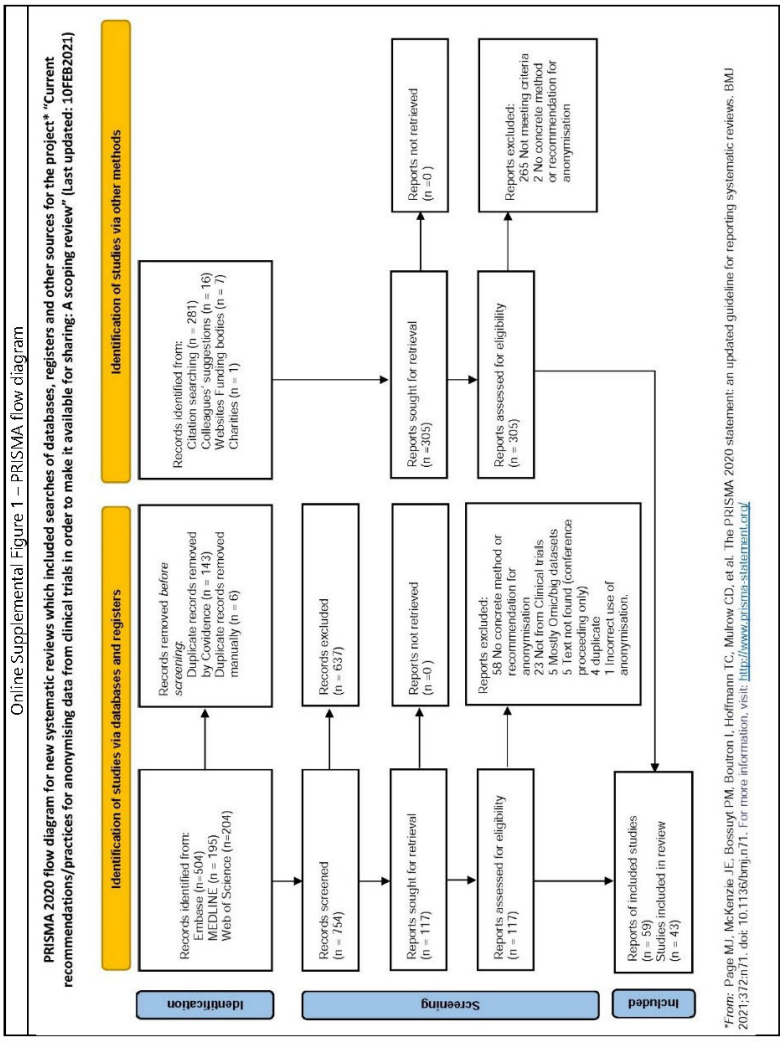
- If one or more answers are YES, then risks have been identified and prior to commencing any data collection **formal ethical review is required** - either:
  - by NHS REC (NB a copy of ethics application and decision letter, and this level 1 form, must be sent to Usher Ethics usher.ethics@ed.ac.uk); or
  - if not to be formally reviewed by NHS REC, then Usher level 2/3 ethical review required.
    [If either 4 is 'yes' or 3b is 'vulnerable' then it is possible level 3 review is required.]

Aryelly Rodriguez	Professor Steff Lewis
Student Name	Supervisor Name
[redacted] 12Dec2018	[redacted] 12Dec2018
Student Signature	Supervisor Signature *

*** NOTE to supervisor:** *The Usher Ethics Subgroup will not check this form (the light touch Level 1 form means we have insufficient detail to do so). By counter-signing this check-list as truly warranting all 'No' answers, **you** are taking responsibility, on behalf of Usher and UoE, that the research proposed truly poses no potential ethical risks. Therefore, if there is any doubt on any issue, it would be a wise precaution to mark it as 'uncertain' and contact the Ethics Subgroup as to whether a level 2 form might be required as well. (See Intra Ethics website – URL at top of form)* *28 Jan 2017*

Appendix 2 - Chapter 3 - Published supplementary materials

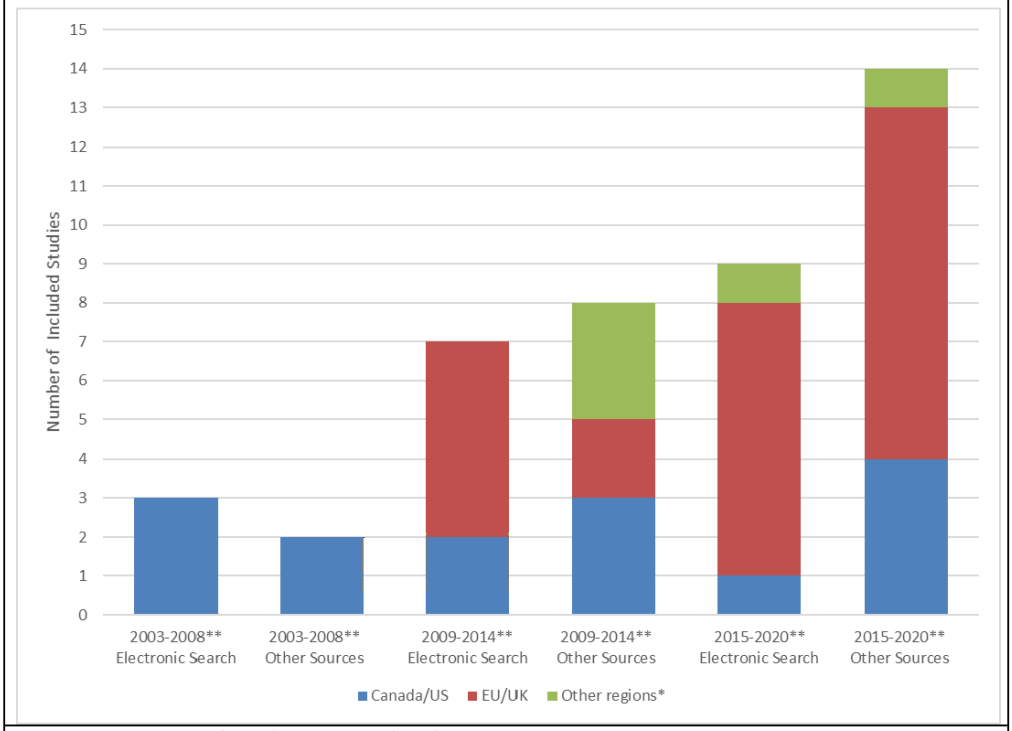
Online supplemental for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review



Final 1.0 09 Feb 2022 Page 1 of 4
File name: Online supplemental RodriguezA_SR_Recomend_Practices_final_01_20220209.docx

Online supplemental for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Online Supplemental Figure 2 – Studies over time by Source and Country/Region



Notes: *Consisting of Australia, US-EU-UK and South Korea.
**Where applicable the oldest record in the included study determined the overall study date.

Online supplemental for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Online Supplemental Figure 3 NVivo® Word Cloud

Notes: Created 01 September 2019.

Online supplemental for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Online Supplemental Table 1 – Identified themes	
Theme description	Theme Id
When authors provided a clear definition of anonymisation, de-identification and pseudonymisation.	Text coded to 1, 2 or 3 respectively
If authors mentioned and described the removal of "The Health Insurance Portability and Accountability Act of 1996" (HIPAA) or Hrynaszkiewics identifiers.	Text coded to 2.1 or 2.2 respectively
If authors described manipulation of data in general.	Text coded to 4
If authors explained with further detail the manipulation of data used (e.g. Perturbation, Recalculation, Recoding, Suppression or Remove superfluous data).	Text coded to 4.1 to 4.5, categories were added as needed
If authors mentioned and described a privacy model [88-90] (i.e. the dataset must satisfy certain conditions to keep the re-identification risk at or below an acceptable level. Privacy models usually depend on an algorithm that uses variables on the dataset to determine how much re-identification risk is present.).	Text coded to 4
If authors explained with further detail the privacy model used (e.g. k-anonymity,[88] l-diversity,[89] differential privacy.[90]).	Text coded to 5.1, categories were added as needed
If authors explained, advocated or provided examples of "controlled access" to anonymised/de-identified datasets (e.g. the use of data sharing agreements, the location of data behind a secure access barrier, the identification and vetoing of secondary researchers (e.g. checking requesters are bona fide researchers with a valid research question).	Text coded to 6
If authors explained with further detail the control access method (e.g. black box where data cannot be seen but the variables are available to make queries and results can be generated, end to end encryption, use of safe haven, or split location for datasets).	Text coded to 6.1 to 6.4, categories were added as needed
If authors explained, advocated or provided examples of "open access" to datasets, where minimal amount of (or non-existent) requirements for allowing access to the data set by secondary researchers. (e.g. free access to the dataset via a website.[91]	Text coded to 7
If authors explained, advocated or provided examples of central repository (e.g. a web page) in which researchers could deposit their dataset regardless of their affiliation.	Text coded to 8
If authors explained that data could be released using expert determination in which a qualified person could deem the risk of re-identification to be small enough to allow such release.	Text coded to 9
If authors explained that other information/documents should be provided with the clinical trial dataset (e.g. annotated case report forms (CRFs), statistical analysis plans (SAPs), study protocols, data dictionaries, anonymisation/de-identification techniques used for generating the released dataset, Clinical Summary Report (CSR), data sharing plans).	Text coded to 10
If authors explained that a re-identification risk assessment should be carried out on the anonymised/de-identified datasets.	Text coded to 11

Appendix 1 - Protocol

Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: Protocol for a scoping review

Rodriguez A[1], Tuck C.[1], Dozier MF[2], Lewis SC[1], Eldridge S[3], Weir CJ[1]

[1]Edinburgh Clinical Trials Unit, Usher Institute of Population Health Sciences and Informatics, the University of Edinburgh

[2]Library & University Collections, Information Services, the University of Edinburgh

[3]Pragmatic Clinical Trials Unit, Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London

Correspondence:

Ms Aryelly Rodriguez

Edinburgh Clinical Trials Unit, the University of Edinburgh

Level 2, Nine Edinburgh BioQuarter,

9 Little France Road, Edinburgh, EH16 4UX

Emails:

Protocol registration

The protocol will not be registered with the International Prospective Register of Systematic Reviews (PROSPERO) as the proposed systematic review does not meet the PROSPERO inclusion criteria.

Appendix 1 - Protocol

Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: Protocol for a scoping review

Abstract

There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. There are various sets of recommendations on how to perform anonymisation prior to sharing clinical trial data. We aim to systematically identify, describe and synthesise these recommendations. We will systematically search literature databases and websites of key organisations in the field. Any publication reporting recommendations on anonymisation to enable data sharing in clinical trials will be included. Two reviewers will independently screen titles, abstracts and full text for eligibility. One reviewer will extract data from included papers which will then be sense checked by a second reviewer. Results will be summarised by narrative review. This scoping review will provide information about existing recommendations for anonymising clinical trial datasets in order to make them available for sharing and it will inform (if applicable) the development of new recommendations.

Key Words: Clinical Trials | Systematic Review | Data Anonymization | Patient Identification Systems | Personally Identifiable Information | Datasets | Data Curation | Guidelines

Appendix 1 - Protocol

Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: Protocol for a scoping review

Background

When academic-led clinical trials are completed, their results are usually released to the public and wider scientific community in scientific journals or clinical trials registries. However, generally, there are considerable amounts of data that are not analysed as part of the final report [1]. In addition, it is often useful to perform meta-analyses across several trials and using the individual patient data from each trial adds to the quality of such analyses [2], for instance by allowing full investigation of subgroup effects. Clinical trials are complex, time-consuming and costly, and it is wasteful not to use data fully [3]. There is now a drive, particularly from publishers and funders, to encourage the release of relevant anonymised trial data sets [4].

Clinical trial datasets contain personal health information on the trial participants. It is imperative that data sharing does not disclose personal data to anyone who falls outside the original group to whom the trial participants consented to disclose their data. Anonymising the trial dataset fulfils this requirement. However, the anonymisation process removes information from the data, and if not done properly, the original trial analyses could not be reproduced, which in turn will limit the data's usability for further research [5]. Anonymisation is complex, and there are many possible ways of performing it.

The drive to share data more widely has generated various sets of recommendations to enable sharing [4, 6-9]. Embedded within these, there is a variety of recommendations on how to anonymise a dataset.

Why it is important to do this review

To our knowledge, there are no reviews of the methods and/or recommendations for the process of generating anonymised clinical trial datasets[1]. To understand and bring together

[1] A quick search was executed on the 07JAN2019 on Google Scholar with "literature" "review" "anonymization" "methods" "clinical trials" and also "literature" "review" "anonymisation" "methods" "clinical trials", the first 100 results were screened for each search and relevant results were not found

the techniques used or recommended for data anonymisation in clinical trials, a systematic review is required.

Objective

To identify, describe and synthesise the existing methods/recommendations to anonymise datasets from clinical trials.

Methods

The Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) guidance [10] and the Joanna Briggs Institute Reviewers' Manual: 2015 edition/ supplement/ Methodology for JBI Scoping Reviews [11] were followed for the development of the protocol, where relevant to a methodology systematic review such as this.

Types of publications

We will include publications giving recommendations on anonymising datasets from clinical trials in any therapeutic area. Non-empirical publications such as editorials, expert views or practice guidelines will also be included in this review

Type of Outcomes

The primary outcome is the reported methods and/or recommendations for anonymisation of clinical trials datasets.

Search methods for identification of publications

We will perform a comprehensive systematic search to identify publications reporting methods or recommendations for anonymising clinical trials datasets. No language restrictions will be used. Non-English publications will be initially translated into English using Google

Translate [12]. If they seem relevant, they will be fully translated as local expertise and resources allow, and any literature that we are unable to translate will be declared. The time period of the execution of the searches will be reported

Electronic Searches

Web of Science, Medline (including non-indexed and in-process records), and Embase databases will be searched from inception to the present day.

The search strategy will use the following key concept areas, adopting subject headings and keywords as relevant for each database:

(Clinical) and (trial* or randomi* or research* or control*) and (principle* or guid* or recomm*) and (shar* or reus* or re-us* or access* or open) and (de-identi* or deidenti* or anonym* or privacy or confidential*)

An example of a detailed electronic search strategy is presented in Appendix 1

Searching other resources

In order to ensure the comprehensiveness of our search, the websites of major

- Research governance organisations,
- Public research funding bodies and charities, and
- Academic clinical trials units

will be searched to find guidelines published as grey literature, so as not to omit documents not published as journal articles and not indexed in the bibliographic databases.

To further supplement our search yield, we will use backwards and forward citation searching on the retrieved documents in order to find additional sources. Also, experts and authors known to have published relevant work will be contacted to identify further literature.

Appendix 1 - Protocol

Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: Protocol for a scoping review

Data collection and analysis

Selection of publications

Records will be retrieved and transferred into a reference manager (e.g. EndNote [13] or Citavi [14]), which will be used for de-duplication and to maintain a master library of the records throughout the review process. Two reviewers (AR, CT) will independently screen titles and abstracts for eligibility. Full text copies of all potentially relevant records will be obtained. Two reviewers (AR, CT) will independently assess whether each full text record meets the inclusion criteria. Covidence software [15] will be used for screening. Any discrepancies will be discussed between the reviewers and if agreement cannot be reached then it will be arbitrated by a third reviewer (SCL, CJW or SE)

Publications will be excluded if:

1. They do not have concrete recommendations/methods of anonymisation.
2. Are not from a clinical trial framework
3. Are focused on omics data or big data

Data extraction and management

A data extraction form to collect relevant data items from eligible sources will be developed and piloted in line with Cochrane guidance [16], this will include (but it is not limited to):

- Publication details (Authors names, Journal, year)
- Publication context (Country, main therapeutic area (if applicable))
- All listed recommendations/methods on anonymisation.

Data extraction will be undertaken by one reviewer (AR) who will manually extract relevant data from each included publication onto the data extraction form, which will be sense checked independently by a second reviewer (CT). Any discrepancies will discussed between the reviewers and if agreement cannot be reach then it will be resolved by a third reviewer

(SCL, CJW or SE). Data extracted will be managed and coded using a qualitative research software (e.g. NVivo® [17] or Citavi [14]).

Data synthesis

Data from the included publication will be summarised in descriptive tables.

Results will be summarised by narrative review and if applicable descriptive statistics.

Conclusions

Currently there is a strong demand for academic researchers to share their research data more readily. In clinical trials, data can be shared more widely if they are anonymised, yet we do not have standardised recommendations on how to do this. To the best of our knowledge, this will be the first systematic review of these emerging recommendations/techniques. We will gather and describe all the identified recommendations and we will provide a map for future research regarding anonymisation of datasets in clinical trials to enable data sharing.

Ethics and dissemination

This project will not collect any patient data or outcomes; therefore, it will not be necessary to seek formal National Health Service Research Ethics Committee's approval. However, we will apply for ethical approval from the Internal Ethics Review Board at The University of Edinburgh's Usher Institute. Findings from the review will be presented at scientific conferences and if possible, will be published in a peer-reviewed journal.

156 **Conflicts of Interests**

157 The authors declare no competing interests.

158

159 **Funding**

160 AR has a scholarship from the University of Edinburgh to undertake a PhD with the support
161 from the Asthma UK Centre for Applied Research (AUKCAR). Neither funder (University of
162 Edinburgh) nor sponsor (AUKCAR) contributed to protocol development. CJW is supported in
163 this work by NHS Lothian via the Edinburgh Clinical Trials Unit.
164 SCL and CT are supported in this work by their employment at the Edinburgh Clinical Trials
165 Unit.
166 SE is supported in this work by her employment at the Pragmatic Clinical Trials Unit.
167 MFD is supported in this work by her employment at the University of Edinburgh.

168

169 **Author contributions**

170 AR, SCL and CJW conceived the idea for this work supported by SE, MFD and TC. AR wrote
171 the first draft, and all authors contributed to the article.

172

Appendix 1 - Protocol

Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: Protocol for a scoping review

Appendix 1. Search Strategy

Medline (Ovid)

1. (clinical adj2 (trial* or randomi* or research* or control*)).mp.

2. (principle* or guid* or recomm*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

3. (shar* or reus* or re-us* or access* or open).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

4. Data Anonymization/

5. (de-identi* or deidenti* or anonym* or privacy or confidential*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]

6. 4 or 5

7. 1 and 2 and 3 and 6

Reference List

1. Song, F., L. Hooper, and Y. Loke, *Publication bias: what is it? How do we measure it? How do we avoid it?* Open Access Journal of Clinical Trials, 2013. **2013**(5): p. 71-81.
2. Berlin, J.A., et al., *Bumps and bridges on the road to responsible sharing of clinical trial data.* Clinical Trials, 2014. **11**(1): p. 7-12.
3. Chan, A.-W., et al., *Increasing value and reducing waste: addressing inaccessible research.* The Lancet, 2014. **383**(9913): p. 257-266.
4. Ohmann, C., et al., *Sharing and reuse of individual participant data from clinical trials: principles and recommendations.* BMJ Open, 2017. **7**(12).
5. El Emam, K. and L. Arbuckle, *Anonymizing health data : case studies and methods to get you started.* 2013: O'Reilly Media, Inc.
6. Hrynaszkiewicz, I., et al., *Preparing raw clinical data for publication trials.* Trials, 2010 **11:9**.
7. Information Commissioner's Office (ICO), *Anonymisation code of practice.* 2012: UK.
8. Keerie, C., et al., *Data sharing in clinical trials - practical guidance on anonymising trial datasets.* Trials, 2018. **19**(1): p. 25.
9. Tudur Smith, C., et al., *Good Practice Principles for Sharing Individual Participant Data from Publicly Funded Clinical Trials.* 2015: UK.
10. Moher, D.S., L.; Clarke, M.; Ghersi, D.; Liberati, A.; Petticrew, M.; Shekelle, P.; Stewart, L.;, *Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement.* Syst Rev, 2015. **4**(1).
11. The Joanna Briggs Institute, *Joanna Briggs Institute Reviewers' Manual: 2015 edition / Supplement*, in *Methodology for JBI Scoping Reviews.* 2015, The Joanna Briggs Institute.
12. Google. *Google Translate.* Available from: https://translate.google.com/.
13. Clarivate Analytics. *EndNote.* Available from: https://endnote.com/.
14. Swiss Academic Software. *CITAVI.* Available from: https://www.citavi.com/en.
15. Covidence. *Covidence.* Available from: https://www.covidence.org/home.
16. Higgins, J.P.T. and S. Green, *Cochrane Handbook for Systematic Reviews of Interventions.* 2011, The Cochrane Collaboration.
17. QSR International. *NVIVO.* Available from: https://www.qsrinternational.com/nvivo/home.

Appendix 2 for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Appendix 2 – Search strategies for electronic databases/registries

11/02/2021

Study Recommendations/methods on anonymisation in clinical trials

Catalogue	Strategy
Ovid MEDLINE(R) and In-Process & Other Non-Indexed Citations Notes: lines 8 and 9 were added to update the search	1.- (clinical adj2 (trial* or randomi* or research* or control*)).mp. 2.- (principle* or guid* or recomm*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 3.- (shar* or reus* or re-us* or access* or open).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 4.- Data Anonymization/ 5.- (de-identi* or deidenti* or anonym* or privacy or confidential*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms] 6.- 4 or 5 7.- 1 and 2 and 3 and 6 8.-(2019* or 2020* or 2021*).ed. 9.- 7 and 8
Ovid Embase Classic + Embase Notes: lines 12 and 13 were added to update the search	1.- (clinical adj2 (trial* or randomi* or research* or control*)).mp. 2.- (principle* or guid* or recomm*).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] 3.- (shar* or reus* or re-us* or access* or open).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] 4.- Data Anonymization/ 5.- (de-identi* or deidenti* or anonym* or privacy or confidential*).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word] 6.- 4 or 5 7.- 1 and 2 and 3 and 6 8.- anonymization/ 9.- 5 or 8 10.- 1 and 2 and 3 and 9 11.- 7 and 10 12.- (2019* or 2020* or 2021*).em. 13.- 11 and 12
Web of Science	# 9 #8 AND #5 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 8 LD=(1900-01-01/2019-02-10) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 7 #6 AND #5 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 6 LD=(2019-02-11/2021-12-31) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 5 #1 AND #2 AND #3 AND #4 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 4 TS=(de-identi* or deidenti* or anonym* or privacy or confidential*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 3 TS=(shar* or reus* or re-us* or access* or open) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 2 TS=(principle* OR guid* OR recomm*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years # 1 TS=(clinical NEAR/2 (trial* OR randomi* OR research* or control*)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

Appendix 3 for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Appendix 3 – List of Included Records

16/01/2021 05:12

Study Recommendations/methods on anonymisation in clinical trials

Study id	Main Study name	Articles in Study	Nodes coded	Article classification	First Authors	Country	Article Complete title	year
1	1g_ANMRD	1	7	Other Sources	Australian National Medical Research Data Storage Facility[1]	Australia	Anonymisation	2016
2	1g_ASTHMA UK	1	6	Other Sources	Asthma UK Centre for Applied Research[2]	UK	ASTHMA UK policy data sharing - Introduction to sharing individual participant data	2015
3	1g_E Emam	1	11	Other Sources	El Emam, Khaled[3]	Canada	Concepts And methods for de-identifying clinical trial data	2014
4	1g_Ebner	1	7	Other Sources	Ebner, Hubert[4]	EU	Piloting the European unified patient identity management (EUPID) concept to facilitate secondary use of neuroblastoma data from Clinical Trials and Biobanking	2016
5	1g_EMA	1	13	Other Sources	European Medicines Agency (EMA)[5]	EU	Data anonymisation - a key enabler for clinical data sharing Workshop report	2018
6	1g_healthdata	1	7	Other Sources	The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation[6]	Canada	Accessing Health and Health-Related Data in Canada	2015
7	1g_hhs	3	4	Other Sources	Food Drug Administration[7]	US	HSS - Availability of masked and de-identified non-summary safety and efficacy data; request for comments	2013
			5	Other Sources	National Institutes of Health (NIH)[8]	US	HSS - Clinical research and the HIPAA privacy rule	2012
			12	Other Sources	U.S. Department of Health & Human Services (HHS)[9]	US	HSS - Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule	2012
8	1g_Hollis	1	7	Other Sources	Hollis, Sally[10]	EU	Best practice for analysis of shared clinical trial data	2016
9	1g_Hughes	1	9	Other Sources	Hughes, Sara[11]	UK	Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach	2014
10	1g_Huser	1	8	Other Sources	Huser, Vojtech[12]	US	Data sharing platforms for de-identified data from human clinical trials	2018
11	1g_IoM	1	21	Other Sources	IOM (Institute of Medicine)[13]	US	Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk	2015
12	1g_IPPC	1	9	Other Sources	International Pharmaceutical Privacy Consortium[14]	US-EU-UK	IPPC White Paper on Anonymisation of Clinical Trial Datasets.	2014
13	1g_Jonas	1	3	Other Sources	Jonas, Stephan[15]	EU	Privacy-Preserving Record Grouping and Consent Management Based on a Public-Private Key Signature Scheme: Theoretical Analysis and Feasibility Study	2019
14	1g_Miller	1	4	Other Sources	Miller, James D[16]	US	Sharing clinical research data in the United States under the health insurance portability and accountability act and the privacy rule	2010
15	1g_MRC	1	5	Other Sources	Medical Research Council (MRC),[17]	UK	GDPR Guidance note 5: Identifiability, anonymisation and pseudonymisation	2019
16	1g_Nelson	1	13	Other Sources	Nelson, Gregory S.[18]	US	Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification	2015
17	1g_NIH	1	6	Other Sources	National Institutes of Health (NIH)[19]	US	NIH data sharing policy and implementation guidance	2003
18	1g_Pfizer	1	5	Other Sources	Pfizer[20]	EU	Clinical Trial Data Access -Policy Document 01312014	2014
19	1g_PhUSE	10	10	Other Sources	Ferran, Jean-Marc[21]	EU	PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models	2015
			9	Other Sources	Ferran, Jean-Marc[22]	EU	PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models - DS10 - Old version of DH01	2015
			4	Other Sources	Ferran, Jean-Marc[23]	EU	PhUSE - De-Identification Standards for CDISC Data Models - PhUSE, Data Transparency Working Group Lead	2017
			6	Other Sources	Iversen, Jørgen Mangor[24]	EU	PhUSE - Data De-Identification Made Simple	2016
			11	Other Sources	Kniola, Lukasz[25]	EU	PhUSE - Data Anonymisation and Risk Assessment Automation	2020
			7	Other Sources	Lyathakula, Santhosh[26]	EU	PhUSE - Data Anonymization Providing clinical trial data to outside researchers	2015
			7	Other Sources	Meeh, Sherry[27]	EU	PhUSE Data De-identification Standard for CDSIC SDTM IG 3.2, and EMA Policy 0070	2016
			4	Other Sources	Meeh, Sherry[28]	EU	PhUSE Data De-identification Standard for CDSIC ADaM 2.1 IG 1.0, and Updates for SDTM IG 3.2	2017
			3	Other Sources	PhUSE[29]	EU	PhUSE DeID Standard - SDTM 3.2 - Appendix 1 - Date Offsetting - v1.91[2]	2015
			6	Other Sources	PhUSE[30]	EU	PhUSE Data De-Identification Standard for SDTM 3.2 -appendix 2-low frequencies-v10-19387	2015
20	1g_Shostak	1	10	Other Sources	Shostak, Jack[31]	US	De-Identification of clinical trials data demystified	2006

Appendix 3 for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Appendix 3 – List of Included Records

16/01/2021 05:12

Study id	Main Study name	Articles in Study	Nodes coded	Article classification	First Authors	Country	Article Complete title	
21	1g_sponsor statements	3	7	Other Sources	Clinical Study Data Request (CSDR)[32]	EU	CSDR - Anonymisation of Clinical Trial Datasets	2015
			9	Other Sources	Clinical Study Data Request (CSDR)[33]	EU	CSDR - Anonymisation of Clinical Trial Datasets – Eli Lilly and Company	2015
			8	Other Sources	Clinical Study Data Request (CSDR)[34]	EU	CSDR - Anonymisation of Clinical Trial Datasets - Eisai	2015
22	1g_transcelerate	2	14	Other Sources	TransCelerate BioPharma Inc[35]	US-EU-UK	TransCelerate-Data de-identification and anonymization of individual patient data in clinical studies	2013
			13	Other Sources	TransCelerate BioPharma Inc[36]	US-EU-UK	TransCelerate-Anonymization of Individual Patient Data in Clinical Studies–A Model Approach	2015
23	1g_Walker	1	9	Other Sources	Walker, Neil[37]	UK	All or Nothing: The False Promise of Anonymity	2017
24	2g_ANDS	1	11	Other Sources	Olesen, Sarah[38]	Australia	Publishing and Sharing Sensitive Data	2011
25	Atzor	1	11	Electronic Search	Atzor, S.[39]	EU	Clinical trial data sharing: From principles to practical implementation - An industry model	2014
26	Demotes-Mainard	1	8	Electronic Search	Demotes-Mainard, J.[40]	EU	How the new European data protection regulation affects clinical research and recommendations?	2019
27	El Emam_2008	1	7	Electronic Search	El Emam, K. [41]	Canada	Protecting privacy using k-anonymity	2008
28	El Emam_2015	1	13	Electronic Search	El Emam, K. [42]	Canada	Anonymising and sharing individual patient data	2015
29	Hrynaszkiewicz	1	12	Electronic Search	Hrynaszkiewicz, I.[43]	UK	Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers	2010
30	Keerie	1	12	Electronic Search	Keerie, C[44].	UK	Data sharing in clinical trials - practical guidance on anonymising trial datasets	2018
31	Lee	1	9	Electronic Search	Lee, J.[45]	Korea	Design of a human-centric de-identification framework for utilizing various clinical research data	2018
32	Malin	1	8	Electronic Search	Malin, B.[46]	US	Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research	2010
33	Morse	1	5	Electronic Search	Morse, R[47]	US	Web-browser encryption of personal health information	2011
34	Nasseh	1	5	Electronic Search	Nasseh, D[48]	EU	Matching study to registry data: maintaining data privacy in a study on family based colorectal cancer	2014
35	Nitzlnader	1	7	Electronic Search	Nitzlnader, M[49]	EU	Patient identity management for secondary use of biomedical research data in a distributed computing environment	2014
36	Noumeir	1	7	Electronic Search	Noumeir, R.[50]	Canada	Pseudonymization of radiology data for research purposes	2007
37	Ohmann	2	8	Electronic Search	Ohmann, C[51]	EU	Sharing and reuse of individual participant data from clinical trials: principles and recommendations	2017
			6	Electronic Search	Ohmann, C - Supplement[51]	EU	Sharing and reuse of individual participant data from clinical trials: principles and recommendations	2017
38	Schell	1	10	Electronic Search	Schell, S[52]	US	Creation of clinical research databases in the 21st century: a practical algorithm for HIPAA Compliance	2006
39	Sudlow	1	7	Electronic Search	Sudlow, R.[53]	UK	EFSPI/PSI working group on data sharing: accessing and working with pharmaceutical clinical trial patient level datasets–a primer for academic researchers	2016
40	Tuck	2	4	Electronic Search	Tuck, C[54]	UK	Data sharing in clinical trials - practical guidance on anonymising trial datasets - Oral Presentation	2015
			8	Electronic Search	Tuck, C[54]	UK	Presentation	2015
41	Tucker	1	15	Electronic Search	Tucker, K.[55]	UK	Protecting patient privacy when sharing patient-level data from clinical trials	2016
42	Tudur-Smith 2015-2017	3	11	Electronic Search	Tudur Smith, C[56]	UK	How should individual participant data (IPD) from publicly funded clinical trials be shared?	2015
			5	Electronic Search	Tudur Smith, C[57]	UK	Good practice principles for sharing individual participant data from publicly funded clinical trials	2015
			8	Electronic Search	Tudur Smith, C[58]	UK	Resource implications of preparing individual participant data from a clinical trial to share with external researchers	2017
43	Wallace	1	3	Electronic Search	Wallace, S.[59]	UK	Protecting personal data in epidemiological research: DataSHIELD and UK law	2014

Appendix 3 for current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review

Reference List

1. Australian National Medical Research Data Storage Facility. Anonymisation. circa 2016.
2. Asthma UK Centre for Applied Research. ASTHMA UK policy data sharing - Introduction to sharing individual participant data. Version 2 ed. circa 2015, p. 2.
3. El Emam K and Malin B. Concepts And methods for de-identifying clinical trial data. *Paper commissioned by the Committee on Strategies for Responsible Sharing of Clinical Trial Data*. 2014.
4. Ebner H, Hayn D, Falgenhauer M, et al. Piloting the European unified patient identity management (EUPID) concept to facilitate secondary use of neuroblastoma data from Clinical Trials and Biobanking. *Health Informatics Meets EHealth: Predictive Modeling in Healthcare–From Prediction to Prevention Proceedings of the 10th EHealth2016 Conference*. IOS Press, 2016, p. 31-8.
5. European Medicines Agency (EMA). Data anonymisation - a key enabler for clinical data sharing Workshop report. In: Agency EM, (ed.). *EMA/796532/2018*. London2018.
6. The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation. Accessing Health and Health-Related Data in Canada. Printed in Ottawa, Canada: Council of Canadian Academies Ottawa, Ontario, Canada, 2015.
7. Food Drug Administration. HHS - Availability of masked and de-identified non-summary safety and efficacy data; request for comments. *Federal Register Available: https://www.federalregister.gov/articles/2013/06/04/2013-13083/availability-of-masked-and-de-identified-non-summary-safety-and-efficacy-data-request-for-comments*. 2013.
8. National Institutes of Health (NIH). HHS - Clinical research and the HIPAA privacy rule. *Retreived from http://privacyruleandresearch.nih.gov/clin_research.asp*. 2004.
9. U.S. Department of Health & Human Services (HHS). HHS - Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. *US Department of Health and Human Services, Washington, DC) Available at :https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html*. 2012; 26.
10. Hollis S, Fletcher C, Lynn F, et al. Best practice for analysis of shared clinical trial data. *BMC medical research methodology*. 2016; 16 Suppl 1: 76.
11. Hughes S, Wells K, McSorley P and Freeman A. Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach. *Pharmaceutical Statistics*. 2014; 13: 179-83.
12. Huser V and Shmueli-Blumberg D. Data sharing platforms for de-identified data from human clinical trials. *Clinical Trials*. 2018; 15: 413-23.
13. IOM (Institute of Medicine). *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington (DC): National Academies Press, 2015.
14. International Pharmaceutical Privacy Consortium. IPPC White Paper on Anonymisation of Clinical Trial Datasets. *Online: http://pharmaprivacy.org/activities/ippc-white-paper-on-anonymisation-of-clinical-trial-data-sets*. 2014.
15. Jonas S, Siewert S and Spreckelsen C. Privacy-Preserving Record Grouping and Consent Management Based on a Public-Private Key Signature Scheme: Theoretical Analysis and Feasibility Study. *Journal of medical Internet research*. 2019; 21: e12300.
16. Miller JD. Sharing clinical research data in the United States under the health insurance portability and accountability act and the privacy rule. *Trials*. 2010; 11: 112.
17. Medical Research Council (MRC). GDPR Guidance note 5: Identifiability, anonymisation and pseudonymisation. 2019.
18. Nelson GS. Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. In: SAS, (ed.). *SAS GLOBAL FORUM Proceedings 2015*. 2015, p. 1-23.
19. National Institutes of Health (NIH). NIH data sharing policy and implementation guidance. *Retrieved June*. 2003; 18: 2009.
20. Pfizer. Clinical Trial Data Access -Policy Document 01312014. 2014.
21. Ferran J-M, El Emam K, Nolan S, Grimm B and De Donder N. PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models. *PhUSE*. 2015.
22. Ferran J-M and Lanoue J. PhUSE De-Identification Working Group: Providing De-Identification Standards to CDISC Data Models - DS10 - Old version of DH01. *PharmaSUG 2015 - Paper DS10 - Old version of DH01*. 2015.
23. Ferran J-M. PhUSE - De-Identification Standards for CDISC Data Models - PhUSE, Data Transparency Working Group Lead. *4th International Clinical Trials Methodology Conference (ICTMC)* Liverpool2017.
24. Iversen JM. PhUSE - Data De-Identification Made Simple. *PHUSE - LEO Pharma A/S, Ballerup, Denmark*. 2016.
25. Kniola L, Hughes A, Paczewska-Sosnowska A, et al. PhUSE - Data Anonymisation and Risk Assessment Automation. *PhUSE*. 2020; 1.0.
26. Lyathakula S. PhUSE - Data Anonymization Providing clinical trial data to outside researchers. In: NOVARTIS, (ed.). *PhUSE Single Day Event (SDE)* Mumbai 2015.
27. Meeh S. PhUSE Data De-identification Standard for CDSIC SDTM IG 3.2, and EMA Policy 0070. In: Janssen IDAaR-, (ed.). . 2016.
28. Meeh S. PhUSE Data De-identification Standard for CDSIC ADaM 2.1 IG 1.0, and Updates for SDTM IG 3.2. 2017.
29. PhUSE. PhUSE DelD Standard - SDTM 3.2 - Appendix 1 - Date Offsetting - v1.91[2]. In: PhUSE, (ed.). PhUSE 2015.
30. PhUSE. PhUSE Data De-Identification Standard for SDTM 3.2 -appendix 2-low frequencies-v10-19387. 2015.
31. Shostak J. De-Identification of clinical trials data demystified. *SAS Users Group*. 2006.
32. Clinical Study Data Request (CSDR). CSDR - Anonymisation of Clinical Trial Datasets. circa 2015.
33. Clinical Study Data Request (CSDR) EL. CSDR - Anonymisation of Clinical Trial Datasets – Eli Lilly and Company. circa 2015.
34. Clinical Study Data Request (CSDR) and Eisai. CSDR - Anonymisation of Clinical Trial Datasets - Eisai circa 2015.
35. TransCelerate BioPharma Inc. TransCelerate-Anonymization of Individual Patient Data in Clinical Studies–A Model Approach. De-identification, Transcelerate-Data, 2015.
36. TransCelerate BioPharma Inc. TransCelerate-Data de-identification and anonymization of individual patient data in clinical studies. *TransCelerate - Clinical Data Transparency Initiative*. 2016.
37. Walker N. All or Nothing: The False Promise of Anonymity. *Data Science Journal*. 2017; 16.
38. Olesen S. Publishing and Sharing Sensitive Data. *Australian National Data Service*. 2011.
39. Atzor S, Sorof J, Kelman A, et al. Clinical trial data sharing: From principles to practical implementation - An industry model. *Regulatory Rapporteur*. 2014; 11: 4-7.
40. Demotes-Mainard J, Cornu C, Guerin A, et al. How the new European data protection regulation affects clinical research and recommendations? *Therapie*. 2019.
41. El Emam K and Dankar FK. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*. 2008; 15: 627-37.
42. El Emam K, Rodgers S and Malin B. Anonymising and sharing individual patient data. *BMJ*. 2015; 350: h1139.
43. Hrynaszkiewicz I, Norton ML, Vickers AJ and Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials*. 2010; 11.
44. Keerie C, Tuck C, Milne G, Eldridge S, Wright N and Lewis SC. Data sharing in clinical trials - practical guidance on anonymising trial datasets. *Trials*. 2018; 19: 25.
45. Lee J, Jung J, Park P, Chung S and Cha H. Design of a human-centric de-identification framework for utilizing various clinical research data. *Human-Centric Computing and Information Sciences*. 2018; 8.
46. Malin B, Karp D and Scheuermann RH. Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research. *Journal of Investigative Medicine*. 2010; 58: 11-8.
47. Morse RE, Nadkarni P, Schoenfeld DA and Finkelstein DM. Web-browser encryption of personal health information. *BMC medical informatics and decision making*. 2011; 11: 70.
48. Nasseh D, Engel J, Mansmann U, Tretter W and Stausberg J. Matching study to registry data: maintaining data privacy in a study on family based colorectal cancer. *e-Health – For Continuity of Care*. 2014.
49. Nitzlnader M and Schreier G. Patient identity management for secondary use of biomedical research data in a distributed computing environment. *eHealth*. 2014, p. 211-8.
50. Noumeir R, Lemay A and Lina JM. Pseudonymization of radiology data for research purposes. *Journal of Digital Imaging*. 2007; 20: 284-95.
51. Ohmann C, Banzi R, Canham S, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open*. 2017; 7.
52. Schell SR. Creation of clinical research databases in the 21st century: a practical algorithm for HIPAA Compliance. *Surgical Infections*. 2006; 7: 37-44.
53. Sudlow R, Branson J, Friede T, Morgan D and Whately-Smith C. EFSPI/PSI working group on data sharing: accessing and working with pharmaceutical clinical trial patient level datasets--a primer for academic researchers. *BMC medical research methodology*. 2016; 16: 73.
54. Tuck C, Lewis S, Milne G, Eldridge S and Wright N. Data sharing in clinical trials - practical guidance on anonymising trial datasets - Oral Presentation. *Trials*. 2015; 16.
55. Tucker K, Branson J, Dilleen M, et al. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Medical Research Methodology*. 2016; 16 Suppl 1: 77.
56. Tudur Smith C, Hopkins C, Sydes MR, et al. How should individual participant data (IPD) from publicly funded clinical trials be shared? *BMC Medicine*. 2015; 13: 298.
57. Tudur Smith C, Hopkins C, Sydes MR, et al. Good practice principles for sharing individual participant data from publicly funded clinical trials Version 1 ed.: Medical Research Council - Hubs for Trials Methodology Research, 2015.
58. Tudur Smith C, Nevitt S, Appelbe D, et al. Resource implications of preparing individual participant data from a clinical trial to share with external researchers. *Trials*. 2017; 18: 319.
59. Wallace SE, Gaye A, Shoush O and Burton PR. Protecting personal data in epidemiological research: DataSHIELD and UK law. *Public Health Genomics*. 2014; 17: 149-57.

Appendix 3 - Chapter 4 - Submitted supplementary materials for publication

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendix 1 Study protocol

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

Rodriguez A[1], Lewis SC[1], Eldridge S[2], Jackson T[3], Weir CJ[1]

[1]Edinburgh Clinical Trials Unit, Usher Institute, the University of Edinburgh

[2]Pragmatic Clinical Trials Unit, Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London

[3]Asthma UK Centre for Applied Research, Usher Institute, the University of Edinburgh

Correspondence:

Ms Aryelly Rodriguez

Edinburgh Clinical Trials Unit, the University of Edinburgh

Level 2, Nine Edinburgh BioQuarter,

9 Little France Road, Edinburgh, EH16 4UX

Emails:

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

Abstract

There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. Some anonymised datasets are now publicly available for secondary research. However, we do not know if they pose a privacy risk to the involved patients. We have 3 equations that can be used to calculate the re-identification risk scores for an entire anonymised dataset, using information in the anonymised dataset. These equations only generate numbers, and they do not aim to actually re-identify individuals in the datasets. We aim to collect a broad sample of publicly available, anonymised clinical trial datasets to calculate their re-identification risk scores. Step 1: We will contact data holders and request access to their anonymised datasets following the data owners' local procedures. Step 2: Re-identification risk scores will be calculated for each dataset, using the 3 equations. Step 3: We will investigate what characteristics of the datasets are associated with increased or decreased risk score, compare the risk scores and their usability, and discuss our findings. To the best of our knowledge, this will be the first study to use these risk of re-identification scores across a range of clinical trials datasets.

Key Words: Clinical Trials | Data Anonymization | Re-identification | De-identification | Datasets | Risk scores

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

38 **Definitions**

Anonymisation	A data set would be considered anonymised if it has been de-identified and then subsequent data manipulation/steps have been taken to further protect the dataset, for example, if a privacy model has been applied (e.g. k-anonymity) or the link with the original non anonymised dataset has been destroyed and this action cannot be reversed.
De-identification	Removal of all personal health information and all other indirect identifiers which could lead to the identification of an individual. The most common de-identification methods are: 1. HIPPA (US Health Insurance Portability and Accountability Act of 1996) Safe harbour, in which 18 identifiers are removed from the datasets [1] [2] 2. Hrynaszkiewicz et al. [3] proposed an enhanced removal of potential identifiers which are commonly present in clinical trials datasets.
Controlled Access	Datasets that can only be accessed if permission is granted by the data holders via their internal procedures.
Open Access:	Datasets that can be accessed without any or minimal restrictions imposed by the data holders.
Publicly available datasets	Data sets that are discoverable and available for sharing via open or controlled access, this data can be located on central repositories or with individual institutions/researchers
Re-identification risk score	Estimated probability of any given individual being re-identified from an anonymised/de-identified dataset. The re-identification risk score depends on the variables available in the dataset, the number of observations in the dataset and on the strategy used to attack the dataset (prosecutor or journalist scenario).
Prosecutor scenario	If the adversary knows that a target individual (for whom identifiers are known) is in the publicly available dataset (released anonymised and/or de-identified) we are under prosecutor re-identification risk scores. This scenario seeks to identify uniqueness in the records of the publicly available dataset.
Journalist scenario	If the adversary sets out to identify any individual from the publicly available dataset just to prove that it can be done by using another dataset for "matching" with the publicly available dataset, then we are under journalist re-identification risk scores.
Matching dataset	An independently obtained dataset with direct identifier, this dataset also need to contain at least two matching variables to link it with the publicly available dataset.
Metadata	Data that describes characteristics and/or contents of anonymised/de-identified datasets

39

40

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

Background

There is now a strong drive, particularly from publishers and funders, to encourage the release of relevant anonymised trial data sets [4].

Data sharing has become so critical in the area of clinical trials that, for example, new grant applications with funding from Cancer Research UK [5] and the Medical Research Council [6] must contain a concrete data-sharing plan. All clinical trials that began enrolling participants on or after 1 January 2019 must have a data sharing plan in the trial's registration [7].

Also, the International Committee of Medical Journal Editors (ICMJE) is encouraging editors "to give priority to publishing the work of authors who have shared their data" [8]

Therefore, data-sharing has become an essential item to disseminate current research, to enable new investigations and to maximise the scientific endeavour [9] [10] . Currently there are a number of such anonymised datasets made publicly available for secondary research via open or controlled access [11] [12].

Anonymisation of data is complex, and complete anonymisation often means that the detail necessary to fully analyse the data is lost. There is therefore a balance between wanting to de-risk a dataset prior to sharing, against wanting it to be sufficiently detailed to answer valid research questions. We propose to take a set of publicly available datasets, and to calculate the re-identification risk scores using the methods described by El-Emam [13]. We will investigate what characteristics of the datasets are associated with increased or decreased risk scores, interpret all calculated risk scores and assess their usability, and discuss our findings.

Why it is important to do this study?

To our knowledge, there are no studies directly using the proposed methods of calculating the re-identification risk scores across a range of publicly available clinical trial datasets.

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

67

68 **Objective**

69 To calculate and describe the re-identification risk scores of publicly available datasets from
70 clinical trials.

71

72 **Outcomes**

73 For each anonymised/de-identified clinical trial dataset we will calculate:

74 1. Number of indirect identifiers present in the datasets as described by Hrynaszkiewicz
75 et al. [3]
76 2. Re-identification Risk Score A (Ra) = The proportion of records that have a re-
77 identification probability higher than 4 pre-defined thresholds (0.1 0.2 0.3 and 0.4) [13],
78 using all indirect identifiers in the dataset
79 3. Re-identification Risk Score B (Rb) = The Worst case scenario or weakest point in the
80 dataset. The smallest unique group of participants (regarding all indirect identifiers in
81 the dataset) generates the highest risk score for the whole dataset.
82 4. Re-identification Risk Score C (Rc) = The expected value or average risk score across
83 all of the records in the dataset, using all indirect identifiers in the dataset.

84 Each re-identification risk scores (A, B and C) will be estimated under the prosecutor and
85 journalist scenario. For the latter a theoretical matching dataset will be generated. The
86 matching datasets are going to be 15 times bigger than the anonymised/de-identified datasets
87 and they will be tailor-made to contain relevant matching indirect identifiers. For further details
88 regarding the calculation of the re-identification risk scores please see Appendix 1

89

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

Methods

Types of datasets

We will include publicly available anonymised/de-identified datasets from randomised controlled clinical trials (RCTs) on human participants that we can successfully acquire from a range of data repositories and locations described in journals. We will only include publicly available datasets that the data holders deem to be anonymised and/or de-identified – we will not include any that are not de-identified/anonymised and are only shared under controlled access arrangements. We will only include studies where the materials are in English or Spanish due to funding limitation.

Search methods for identification of datasets

Thanks to the results from "Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review" [14] we have identified several data repositories which will be searched and we will seek to obtain as many relevant datasets as possible, through open or controlled access (See Appendix 2).

Clinical trials published in journals with strong data sharing policies (PlosOne and BMJ) will also be searched, from Jan 2013 for BMJ and Mar 2014 for PlosOne (this is when their data sharing policies were introduced) to the present day, in order to locate publicly available datasets. Finally we will not discard any dataset that we find via other sources or in other data repositories not mentioned in appendix 2, if it meets the inclusion criteria, this will be reported.

Data collection and analysis

Selection of datasets

One investigator (AR) will search data repositories and collect all available metadata about the potentially eligible datasets. This will be put on an excel spreadsheet (the datasets' attributes to be collected are shown in Appendix 3).

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

AR will screen titles and/or description of datasets for eligibility and to identify duplicates. Any queries will be raised to SCL and/or CJW.

Datasets will be excluded if:

1. They are not explicitly declared as anonymised/de-identified and suitable for sharing
2. They are not from a RCT
3. They are not from human participants
4. They are in a language that is not English or Spanish

If a single data repository has less or equal than five datasets meeting the inclusion criteria, we would seek to obtain all datasets from that repository. If a repository has more than five eligible datasets, a random selection of five datasets will be drawn from the eligible datasets. This process will be executed and documented by AR. Most data repositories/sources offer the data free of charge, for this project we have modest funding to pay for small nominal fees to the data holders. Unfortunately very expensive request for fee will be declined and reported as part of the results.

Data extraction and management

Selected datasets will be retrieved from data repositories/sources and transferred to AR's secured area at UoE (\\cmvm.datastore.ed.ac.uk\cmvm\smgphs\users\arodrigu) or AR's DataStore allocation as per University of Edinburgh data handling policies [15] [16] [17]. All these datasets will be analysed using SAS 9.4 [18] or more recent version. If datasets are going to be held on the data owners' location, AR will transfer the re-identification risk score calculation analysis SAS 9.4 code to the data holders, in order to perform the analysis at the owners' remote location and she will extract only the relevant output as per the local policies for the remote location. If required by any remote location data holders, the re-identification risk score calculation analysis code could be created on IBM SPSS or R, to perform the analysis, however SAS 9.4 is still the preferred option.

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

142

143 **Data synthesis and re-identification risk calculation**

144 Metadata (Appendix 2) from all included datasets will be summarised in descriptive tables.
145 The number of indirect identifier in the anonymised/de-identified datasets will be done by
146 visual inspection by AR and double checked by a second review (SCL or CJW). We will
147 calculate the several re-identification risk scores with the formulas in chapter 16 from
148 "Guide to the de-identification of personal health information" by Khaled El Eman [13] (for
149 further details on re-identification risk scores calculations see appendix 1) using SAS 9.4.

150

151 Re-identification risk scores from the anonymised/de-identified datasets will be summarised
152 by descriptive statistics. If issues arise with any particular dataset, they will be directly
153 discussed with the datasets owners, and if appropriate, the issue will be reported as a result
154 in this study, anonymised and unlinked to its original anonymised/de-identified dataset. There
155 will not be any attempt to re-identify or contact individual patients.

156

157 **Data reporting and interpretation**

158 All the re-identification risk scores in this protocol aim to assess the level of granularity in a
159 dataset and they complement each other, so we cannot recommend one over the other. The
160 re-identification risk scores are only driven by the number of unique indirect identifiers and the
161 number of records in the dataset.
162 After calculations, we should be able to tell how datasets of a similar size with the same
163 amount of indirect identifiers, compare to each other (e.g. datasets with 10 to 100 patients and
164 2-3 indirect identifiers have risk scores of around 0.3, while datasets with 100-500 patients
165 and 2-3 indirect identifiers have risk scores of 0.2, and datasets with 500 or more patients and
166 2-3 indirect identifiers have risk scores of 0.1). The re-identification risk scores will be
167 generated under both prosecutor and journalist scenarios. To help us explore their meaning,
168 the following plots will be generated (see Table 1):

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

169

Table 1. Plots for Re-identification Risk Scores	
1	Scatterplot of Ra vs anonymised clinical trial datasets' sample size
2	Box-and-whisker plots of Ra vs number of indirect identifiers
3	Scatterplot of Rb vs anonymised clinical trial datasets' sample size
4	Box-and-whisker plots of Rb vs number of indirect identifiers
5	Scatterplot of Rc vs anonymised clinical trial datasets' sample size
6	Box-and-whisker plots of Rc vs number of indirect identifiers
7	Scatterplot of Ra vs Rb
8	Scatterplot of Rb vs Rc
9	Scatterplot of Ra vs Rc
Where: Ra	The proportion of records in the anonymised dataset that have a re-identification probability higher than a priori predetermined threshold.
Rb	The maximum probability of re-identification among all records in the anonymised dataset.
Rc	The proportion of records in the anonymised dataset that could be correctly re-identified on average.
Plots will be done for all collected anonymised clinical trial datasets.	

170

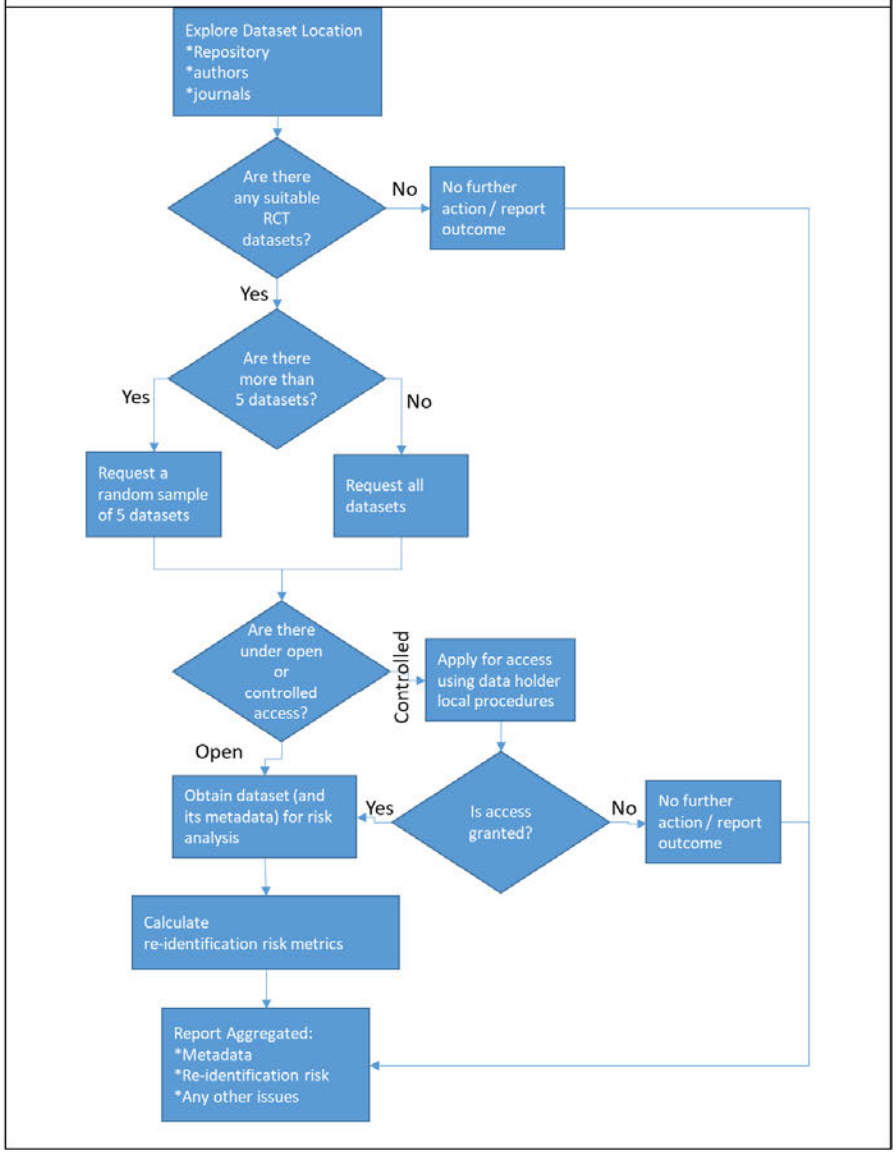
171 The re-identification risk scores by themselves cannot tell if the data has been sufficiently
172 anonymised, but once this project is finalised they could be used to help calibrate the
173 anonymisation process of clinical trials datasets, because they would allow dataset owners to
174 see how much risk other researchers have taken and how theirs compares with that.
175
176 Finally, the re-identification risk scores do not have the capability of determining the probability
177 of re-identification in the real world for a dataset. This is controlled by other factors such as:
178 controlled vs open access to the dataset, attacker's motivations, resources and potential gains
179 and dealing with a stigmatising intervention/disease. We will consider these factors in our final
180 discussion when we get access to the datasets.
181 Figure 1 gives an overview of the process to be followed by this protocol.
182

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

183

Figure 1. Project Flow chart



184

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

185

186 **Potential sources of bias and limitations**

187 This study is covering an emerging part of clinical research, for which even consensus about
188 the definition of anonymisation does not exist. Therefore we will consider datasets defined as
189 anonymised or de-identified. This will create heterogeneity among the characteristics of the
190 datasets that would be requested by this study.
191 We intend to explore the first wave of available datasets, which at the moment, seem to be
192 based mostly in the USA and the UK. Therefore it might not be possible to describe what is
193 happening outside of these countries.
194 We are aware that there are at least 19 open access datasets that we could get a hold of,
195 which means that this project is feasible. However, there is no guarantee that we will be
196 granted access to other datasets.

197

198 **Ethics and dissemination**

199 This project will collect patient data or outcomes from clinical trials datasets that have
200 been anonymised for secondary use; therefore, this is a low risk project and we will follow
201 the ethical review processes coordinated by the Internal Ethics Review Board at The
202 University of Edinburgh's Usher Institute. Findings from this research will be presented
203 at scientific conferences and published in a peer-reviewed journal. No publication or
204 presentation originating from this work will reveal any data that could lead to re-
205 identification of individuals from the data sets used.

206

207 **Conflicts of Interests**

208 The authors declare no competing interests.

209

210

211

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

212 **Funding**

213 AR has a scholarship from the University of Edinburgh to undertake a PhD with the support
214 from the Asthma UK Centre for Applied Research (AUKCAR). Neither funder (University of
215 Edinburgh) nor sponsor (AUKCAR) contributed to protocol development. CJW is supported
216 in this work by NHS Lothian via the Edinburgh Clinical Trials Unit.
217 SCL is supported in this work by her employment at the Edinburgh Clinical Trials Unit.
218 SE is supported in this work by her employment at the Pragmatic Clinical Trials Unit.
219

220 **Author contributions**

221 AR, SCL and CJW conceived the idea for this work supported by SE. AR wrote the first draft,
222 and all authors contributed to this article.
223

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

Appendix 1 Details for Risk Scores Calculation

- **Background**

El-Emam Risk calculation – The equations in table 1 to calculate re-identification risks scores were taken from El Emam K. Guide to the de-identification of personal health information. Chapter 16. CRC Press; 2013 May 6. [13]

Table 1.1. Risk Scores.

Id	Type of Risk Score	Equation for Risk Score	Dichotomous decision rule
Ra	The proportion of records in the anonymised dataset that have a re-identification probability higher than a priori predetermined threshold.	$R_a = \frac{1}{n}(\sum_{j \in J} f_j \times I(\theta_j > \tau))$	$D_a = \begin{cases} HIGH, & R_a > \alpha \\ LOW, & R_a \leq \alpha \end{cases}$
Rb	The maximum probability of re-identification among all records in the anonymised dataset.	$R_b = \max_{j \in J}(\theta_j)$	$D_b = \begin{cases} HIGH, & R_b > \tau \\ LOW, & R_b \leq \tau \end{cases}$
Rc	The proportion of records in the anonymised dataset that could be correctly re-identified on average.	$R_c = \frac{1}{n}(\sum_{j \in J} f_j \theta_j)$	$D_c = \begin{cases} HIGH, & R_c > \lambda \\ LOW, & R_c \leq \lambda \end{cases}$

Where
τ = the highest allowable probability of correctly re-identifying a single record
α= the proportion of records that have a high probability of re-identification that would be acceptable to the data custodian.
λ = the average proportion of records that can be correctly re-identified that would be acceptable to the data custodian.
$I(.)$= the indicator function (this returns 1 if the parameter is true and 0 otherwise).
f_j = the number of individuals in an equivalence class j in the anonymised dataset.
J= the set of equivalence classes in the anonymised dataset.
θ_j= the probability of re-identification of an equivalence class j (all of the records in the same equivalence class will have the same probability value).
n= the total number of records in the anonymised dataset.

The equations in table 1.1 will be adjusted as required to take into consideration if we are under prosecutor or journalist re-identification risk scenarios. Table 1.2 shows how the equations in table 1.1 are adapted to the mentioned re-identification scenarios.

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

234

Table 1.2 Derived Metrics by Risk Scenario

Id	Type Scenario	Equation for Risk
p_Ra	Prosecutor	$p\_R_a = \frac{1}{n}(\sum_{j \in J} f_j \times I(\frac{1}{f_j} > \tau)$
p_Rb		$p\_R_b = \frac{1}{\min_{j \in J}(f_j)} = \max_{j \in J}\left(\frac{1}{f_j}\right)$
p_Rc		$p\_R_c = \frac{1}{n}(\sum_{j \in J} f_j \times \frac{1}{f_j}) = \frac{\lvert J \rvert}{n}$
j_Ra	Journalist*	$j\_R_a = \frac{1}{n}(\sum_{j \in J} f_j \times I(\frac{1}{F_j} > \tau)$
j_Rb		$j\_R_b = \frac{1}{\min_{j \in J}(F_j)} = \max_{j \in J}\left(\frac{1}{F_j}\right)$
j_Rc		$j\_R_c = \max\left(\frac{\lvert J \rvert}{\sum_{j \in J} F_j}, \frac{1}{n}\sum_{j \in J} \frac{f_j}{Fj}\right)$

Where τ = the highest allowable probability of correctly re-identifying a single record.
f_j = the number of individuals in an equivalence class j in the anonymised dataset.
J = the set of equivalence classes in the anonymised dataset.
$|J|$ = the number of unique equivalence classes in the anonymised dataset.
$I(.)$ = the indicator function (this returns 1 if the parameter is true and 0 otherwise).
F_j = the number of individuals in an equivalence class j in the matching dataset
n = the total number of records in the anonymised dataset.
N = the total number of records in the matching dataset.

*These metrics are suitable for the situation where the anonymised dataset is a proper subset of a theoretical or actual matching dataset.

235
236
237

- **Calculated example**

238 This example shows how the calculations will be executed on the datasets that are made
239 available through this protocol. Table 2 displays a mock de-identified 25 observations dataset
240 with two independent indirect identifiers, gender_coded and age_group. The variables
241 gender, year_of_birth and age are not expected to be on actual anonymised datasets (if
242 gender_coded and age_group are present), they are shown for facilitating the explanation of
243 the example.

244

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

245

Table 2. Mock Anonymised/De-identified dataset (25 observations, 11 unique classes)

ID	Gender	Gender_coded	Year_of_Birth	Age	Age_group
1	Male	1	1933	87	1
2	Male	1	1938	82	1
3	Male	1	1948	72	2
4	Female	2	1949	71	2
5	Female	2	1952	68	3
6	Male	1	1955	65	3
7	Male	1	1956	64	3
8	Female	2	1957	63	3
9	Male	1	1960	60	4
10	Female	2	1963	57	4
11	Male	1	1965	55	4
12	Male	1	1965	55	4
13	Female	2	1965	55	4
14	Female	2	1972	48	5
15	Male	1	1973	47	5
16	Male	1	1974	46	5
17	Female	2	1975	45	5
18	Male	1	1976	44	5
19	Female	2	1980	40	6
20	Male	1	1987	33	6
21	Male	1	1982	38	6
22	Female	2	1983	37	6
23	Male	1	1987	33	6
24	Female	2	1981	39	6
25	Male	1	1984	36	6

Unique class (j)	Gender	Age_group (label)	Number of Observation in each Class (fj)
1	Male	80+	1
2	Male	71-80	2
3	Male	61-70	2
4	Male	51-60	3
5	Male	41-50	3
6	Male	30-40	4
--	Female	80+	0
7	Female	71-80	1
8	Female	61-70	2
9	Female	51-60	2
10	Female	41-50	2
11	Female	30-40	3
TOTAL (n)			25

246

247 Prosecutor risk calculation (table 3.1) – We obtained three measures of risk under

248 prosecutor risk for the mock anonymised/de-identified dataset presented in table 2.

249

Table 3.1

Classes in De-identified dataset and interim calculations for Risk Scores

Prosecutor Risk Scores Calculation*

Prosecutor risk tau = 0.33 (Set a priori)

Unique class (j)	Gender coded	Age_group	frequency by class (fj)	Individual class risk (1/fj)	Is (1/fj)>tau? (Ij)	Ij*fj
1	1	1	1	1.0000	TRUE	1
2	1	2	2	0.5000	TRUE	2
3	1	3	2	0.5000	TRUE	2
4	1	4	3	0.3333	TRUE	3
5	1	5	3	0.3333	TRUE	3
6	1	6	4	0.2500	FALSE	0
--	2	1	0		N/A	N/A
7	2	2	1	1.0000	TRUE	1
8	2	3	2	0.5000	TRUE	2
9	2	4	2	0.5000	TRUE	2
10	2	5	2	0.5000	TRUE	2
11	2	6	3	0.3333	TRUE	3
			25			21

R1a	84%	The proportion of records that have a re-identification probability higher than 0.33. (R1a=21/25)
R1b	100%	Worst case scenario. The class with the highest individual risk represents the whole dataset (R1b=max(1/fj)=1)
R1c	44%	Average risk across all of the records in the dataset (Expected value) (R1c=11/25)

*Note that for R1a, a response curve will be generated as several values of tau will be explored

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

250

251 Journalist risk (table 3.2) – We assumed that a matching dataset exists (with 340 observations
252 for this example). We obtained three measures of risk under journalist risk for the mock
253 anonymised/de-identified dataset presented in table 2

254

Table 3.2

Classes in Anonymised/De-identified dataset and interim calculations for Risk Scores

Journalist Risk Scores Calculations*

Journalist risk tau = 0.09 (Set a priori)

Unique class (j)	Gender coded	Age_group	frequency by class (fj)	Fj (in matching dataset)	Individual class risk (fj/Fj)	1/Fj	is (1/Fj)>tau? (Ij)	Ij*fj
1	1	1	1	10	0.1000	0.1000	TRUE	1
2	1	2	2	20	0.1000	0.0500	FALSE	0
3	1	3	2	20	0.1000	0.0500	FALSE	0
4	1	4	3	30	0.1000	0.0333	FALSE	0
5	1	5	3	40	0.0750	0.0250	FALSE	0
6	1	6	4	50	0.0800	0.0200	FALSE	0
..	2	1	0	10	0.0000	0.1000	TRUE	0
7	2	2	1	20	0.0500	0.0500	FALSE	0
8	2	3	2	20	0.1000	0.0500	FALSE	0
9	2	4	2	30	0.0667	0.0333	FALSE	0
10	2	5	2	40	0.0500	0.0250	FALSE	0
11	2	6	3	50	0.0600	0.0200	FALSE	0
			25	**340**	**0.8817**			**1**

R2a	**4%**	The proportion of records that have a re-identification probability higher than 0.09. (R2a=1/25)
R2b	**10%**	Worst case scenario. The class with the highest individual risk represents the whole dataset (R2b=max(1/Fj)=0.1)
R2c1	3.5%	R2c1=0.8817/25
R2c2	3.2%	R2c2=11/340
R2c	**3.5%**	The maximun between the prob of a randomly selected record from the De-identified dataset that is matched against the extra information dataset and the probability of a randomly selected record from the extra information dataset that is matched against the De-identified dataset (R2c=max(3.5,3.2))

*Note that for R2a, a response curve will be generated, as several values of tau will be explored

255

256

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

257

258 ***Appendix 2. Dataset repositories identified during "Current***

259 ***recommendations/practices for anonymising data from clinical trials in order to***

260 ***make it available for sharing: A scoping review" [14]***

261

Id	*Data sharing repository*	*country*	*funding*	*Type Of Access*	*Potential number of studies to check (1)*
1	https://datacompass.lshtm.ac.uk	*UK*	*public*	*Controlled*	*65*
2	https://ctu-app.lshtm.ac.uk/freebird	*UK*	*public*	*Open*	*3*
3	https://datashare.is.ed.ac.uk	*UK*	*public*	*Controlled*	*100*
4	http://datadryad.org	*USA*	*public*	*Open/ controlled (2)*	*180*
5	https://www.clinicalstudydatarequest.com	*UK*	*public and private*	*Controlled*	*56*
6	http://yoda.yale.edu	*USA*	*public and private*	*Controlled*	*38*
7	https://www.projectdatasphere.org	*USA*	*public and private*	*Controlled*	*124*
8	https://biolincc.nhlbi.nih.gov/studies	*USA*	*public*	*Controlled*	*167*
9	https://nda.nih.gov/get/access-data.html	*USA*	*public*	*Controlled*	*25*
10	https://vivli.org/	*USA*	*public and private*	*Controlled*	*17*
11	https://beta.ukdataservice.ac.uk/datacatalogue/studies (reshare.ukdataservice.ac.uk) (https://www.ukdataservice.ac.uk/deposit-data)	*UK*	*public and private*	*Open/ Controlled (2)*	*29*
12	https://med.data.edu.au/find-data/	*Australia*	*public*	*Controlled*	*12*
13	https://dcri.org/our-approach/data-sharing/soar-data SOAR data: Available datasets: Duke cardiac catheterization datasets.	*USA*	*public*	*Controlled*	*4*

262 *(1) Some items are under controlled other seems to be open access*
263 *(2) Web pages visited on the 03FEB2020.*
264

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

265

266 ***Appendix 3 – Proposed Variables in datasets' metadata extraction forms***

267 A data extraction form to collect relevant metadata items from eligible datasets it will include

268 (but it is not limited to):

269

270 • Study name/acronym

271 • Study description

272 • Study design

273 • Study objectives

274 • Study patient population

275 • Study Time frames (e.g trial duration, follow-up duration)

276 • Clinical phase (i.e. I to IV)

277 • Number and location of sites

278 • Date of associated main publication

279 • Therapeutic area

280 • Nature of active Intervention (e.g. investigational medicinal product, surgery or
281 therapy, route of administration)

282 • Nature of control intervention (e.g. placebo, routine care)

283 • Funding source

284 • Sponsor

285 • Documentation availability (e.g. protocol, analysis plan, results, data dictionary)

286

287

288

289

Appendix 1 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol

Reference List

1. Act, A., *Health insurance portability and accountability act of 1996.* Public law, 1996. **104**: p. 191.
2. U.S. Department of Health & Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* 2012.
3. Hrynaszkiewicz, I., et al., *Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers.* Trials, 2010. **11**(340).
4. Dal-Re, R., *Access to Anonymized Individual Participant Clinical Trials Data: A Radical Change of Mind by the Most Prestigious Medical Journals.* Archivos de Bronconeumologia, 2018. **54**(2): p. 65-67.
5. UK, C.R. *Data sharing guidelines.* [cited 2020 30 Oct 2020]; Available from: https://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy/data-sharing-guidelines.
6. MRC, T., *MRC Policy on Open Research Data from Clinical Trials and Public Health Intervention Studies* 2016.
7. Taichman, D.B., et al., *Data sharing statements for clinical trials.* BMJ, 2017. **357**: p. j2372.
8. The EQUATOR Network. *New ICMJE Recommendations published* 2018; Available from: https://www.equator-network.org/2018/12/21/new-icmje-recommendations-published/.
9. Pisani, E., et al., *Beyond open data: realising the health benefits of sharing data.* BMJ, 2016. **355**: p. i5295.
10. Bertagnolli, M., et al., *Advantages of a truly open-access data-sharing model.* N Engl J Med, 2017. **12**(376): p. 1178-1181.
11. (CSDR), C.S.D.R. *Clinical Study Data Request.* Available from: https://clinicalstudydatarequest.com/.
12. University, T.Y. *Yale University Open Data Access (YODA) Project.* [cited 2020 26 Oct 2020]; Available from: http://yoda.yale.edu/.
13. El Emam, K., *Guide to the de-identification of personal health information.* 2013: CRC Press.
14. Rodriguez, A., et al., *Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review.* 2020, University of Edinburgh.
15. Edinburgh, T.U.o. *Working with sensitive data* 2020 [cited 2020 30 Oct 2020]; Available from: https://www.ed.ac.uk/information-services/research-support/research-data-service/during/sensitive-data.
16. Edinburgh, T.U.o. *Use University services.* 2020 [cited 2020 30 Oct 2020]; Available from: https://www.ed.ac.uk/infosec/information-protection-policies/procedures-guidance/use-university-services.
17. Edinburgh, T.U.o. *Data - Data Services.* 2020 [cited 2020 30 Oct 2020].
18. Inc, S.I., *SAS 9.4 [Computer software].* 2013: Cary, NC, USA.

Appendix 2 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendix 2 Dataset repositories

Repositories/data sources from 1-15 were identified during "Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review" [1], repositories from 16-18 were found through alternative sources (web searches or word of mouth)

Id	*Data sharing repository*	Country	Funding	Type of Access	Inclusion
1	https://datacompass.lshtm.ac.uk[2]	UK	public	Controlled	Yes
2	https://ctu-app.lshtm.ac.uk/freebird[3]	UK	public	Open	Yes
3	https://datashare.is.ed.ac.uk[4]	UK	public	Controlled	Yes
4	https://www.clinicalstudydatarequest.com[5]	UK	public and private	Controlled	Yes
5	http://datadryad.org[6]	USA	public	Open	Yes
6	http://yoda.yale.edu[7]	USA	public and private	Controlled	Yes
7	https://www.projectdatasphere.org[8]	USA	public and private	Controlled	Yes
8	https://biolincc.nhlbi.nih.gov/studies[9]	USA	public	Controlled	Yes
9	https://nda.nih.gov/get/access-data.html[10]	USA	public	Controlled	Yes
10	https://vivli.org/[11]	USA	public and private	Controlled	Yes
11	https://beta.ukdataservice.ac.uk/datacatalogue/studies (reshare.ukdataservice.ac.uk) (https://www.ukdataservice.ac.uk/deposit-data) [12]	UK	public and private	Hybrid (1)	Yes
12	https://med.data.edu.au/find-data/[13]	Australia	public	Controlled	No (2)
13	https://dcri.org/our-approach/data-sharing/soar-data SOAR data: Available datasets: Duke cardiac catheterization datasets. [14]	USA	public	Controlled	No (3)
14	https://journals.plos.org/plosone/search [15]	USA	public and private	Hybrid (1)	Yes
15	https://www.bmj.com/search/advanced [16]	UK	private	Hybrid (1)	Yes
16	https://dataverse.harvard.edu/ [17]	USA	public	Open	Yes
17	https://arlg.org/studies-in-progress/ [18]	USA	private	NA	No (3)
18	https://repository.niddk.nih.gov/studies/ [19]	USA	public	Hybrid (1)	No (3)

(1) Some items are under controlled other seems to be open access
(2) Excluded as the repository does not longer exists
(3) Excluded because suitable Randomised Controlled Trials datasets could not be located at the moment of search.

Appendix 2 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Reference List

1. Rodriguez A, Tuck C, Dozier MF, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. *Clinical Trials* 2022; 19: 452-463.
2. London School of Hygiene & Tropical Medicine. LSHTM Data Compass, https://datacompass.lshtm.ac.uk (2020).
3. Clinical Trials Unit London School of Hygiene & Tropical Medicine. The FreeBIRD Bank of Injury and Emergency Research Data, https://freebird.lshtm.ac.uk/ (2020).
4. The University of Edinburgh. Edinburgh DataShare, https://datashare.ed.ac.uk/ (2020).
5. Clinical Study Data Request (CSDR). Clinical Study Data Request, https://clinicalstudydatarequest.com/ (2020, 2020).
6. Dryad. Data Dryad, https://datadryad.org/ (accessed 2020).
7. The Yale University. Yale University Open Data Access (YODA) Project, http://yoda.yale.edu/ (2020, 2020).
8. CEO Roundtable on Cancer Inc. Project Data Sphere, https://www.projectdatasphere.org/ (2020).
9. The National Heart LaBIN. BioLINCC, https://biolincc.nhlbi.nih.gov/ (2020).
10. The National Institute of Mental Health. The NIMH Data Archive (NDA), https://nda.nih.gov/ (2020).
11. Vivli Center for Global Clinical Research Data. Vivli, a global data-sharing and analytics platform. , https://vivli.org/ (2020, 2020).
12. UK Data Service. UK Data Service: data Catalogue, https://beta.ukdataservice.ac.uk/datacatalogue/studies (accessed 2020).
13. Intersect Australia Limited - Queensland Cyber Infrastructure Foundation Ltd. Australian National Medical Research Data Storage Facility, https://med.data.edu.au/find-data/ (accessed 2020).
14. Institute DCR. SOAR DATA™, https://dcri.org/our-approach/data-sharing/soar-data (2020).
15. PLOS is a nonprofit 501(c)(3) corporation. PLOS ONE: An inclusive journal community working together to advance science by making all rigorous research accessible without barriers, https://journals.plos.org/plosone/search.
16. BMJ Publishing Group Ltd. BMJ is a global healthcare knowledge provider with a vision for a healthier world. We share knowledge and expertise to improve healthcare outcomes., https://www.bmj.com/search/advanced.
17. Harvard University. Harvard Dataverse Repository. Deposit and share your data. Get academic credit. Harvard Dataverse is a repository for research data. Deposit data and code here., https://dataverse.harvard.edu/.
18. Antibacterial Resistance Leadership Group (ARLG) ARLG studies, https://arlg.org/summary-of-results/.
19. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). NIDDK Central Repository, https://repository.niddk.nih.gov/studies/dpp/.

Appendix 3 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendix 3 Variables in datasets' metadata

Metadata items from eligible datasets included:

- Study name/acronym
- Study description
- Study design
- Study objectives
- Study patient population/ Therapeutic area
- Study time frames (e.g., trial duration, follow-up duration)
- Study number of participants/patients (sample size)
- Type of access (Open vs Controlled)
- Clinical phase (i.e., I to IV)
- Number and location of sites
- Date of associated main publication
- Country of lead author in main publication
- Primary outcome statistically significant (Yes/No)
- Nature of active Intervention (e.g., investigational medicinal product, surgery or therapy, route of administration)
- Nature of control intervention (e.g., placebo, routine care)
- Funding source
- Sponsor
- Documentation availability (e.g., protocol, analysis plan, results, data dictionary)

Appendix 4 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendix 4 Worked example for Calculating Re-identification risk scores

- **Background**

El-Emam Risk calculation – The equations in table S4.1.1 and S4.1.2 to calculate re-identification risks scores were taken from El Emam K. Guide to the de-identification of personal health information. Chapter 16. CRC Press; 2013 May 6. (El Emam 2013)

Table S4.1.1 shows the general equations to calculation the re-identification risk scores.

Table S4.1.1. Re-identification Risk Scores.

Id	Type of Risk Score	Equation for Risk Score	Dichotomous decision rule
Ra	The proportion of records in the anonymised dataset that have a re-identification probability higher than a priori predetermined threshold.	$R_a = \frac{1}{n}(\sum_{j \in J} f_j \times I(\theta_j > \tau))$	$D_a = \begin{cases} HIGH, & R_a > \alpha \\ LOW, & R_a \leq \alpha \end{cases}$
Rb	The maximum probability of re-identification among all records in the anonymised dataset.	$R_b = \max_{j \in J}(\theta_j)$	$D_b = \begin{cases} HIGH, & R_b > \tau \\ LOW, & R_b \leq \tau \end{cases}$
Rc	The proportion of records in the anonymised dataset that could be correctly re-identified on average.	$R_c = \frac{1}{n}(\sum_{j \in J} f_j \theta_j)$	$D_c = \begin{cases} HIGH, & R_c > \lambda \\ LOW, & R_c \leq \lambda \end{cases}$

Where
τ = the highest allowable probability of correctly re-identifying a single record
α= the proportion of records that have a high probability of re-identification that would be acceptable to the data custodian
λ = the average proportion of records that can be correctly re-identified that would be acceptable to the data custodian
$I(.)$= the indicator function (this returns 1 if the parameter is true and 0 otherwise)
f_j = the number of individuals in an equivalence class j in the anonymised dataset
J= the set of equivalence classes in the anonymised dataset
θ_j= the probability of re-identification of an equivalence class j (all of the records in the same equivalence class will have the same probability value)
n= the total number of records in the anonymised dataset

The equations in table S4.1.1 need to be adjusted to take into consideration if we are under prosecutor or journalist re-identification risk scenarios. Table S4.1.2 shows how the equations in table S4.1.1 are adapted to the mentioned re-identification scenarios.

Appendix 4 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Table S4.1.2 Derived Metrics by Risk Scenario

Id	Type Scenario	Equation for Risk
p_Ra	Prosecutor	$p\_R_a = \frac{1}{n}(\sum_{j\in J} f_j \times I(\frac{1}{f_j} > \tau)$
p_Rb		$p\_R_b = \frac{1}{\min_{j\in J}(f_j)} = \max_{j\in J}\left(\frac{1}{f_j}\right)$
p_Rc		$p\_R_c = \frac{1}{n}(\sum_{j\in J} f_j \times \frac{1}{f_j}) = \frac{\lvert J\rvert}{n}$
j_Ra	Journalist*	$j\_R_a = \frac{1}{n}(\sum_{j\in J} f_j \times I(\frac{1}{F_j} > \tau)$
j_Rb		$j\_R_b = \frac{1}{\min_{j\in J}(F_j)} = \max_{j\in J}\left(\frac{1}{F_j}\right)$
j_Rc		$j\_R_c = \max\left(\frac{\lvert J\rvert}{\sum_{j\in J} F_j}, \frac{1}{n}\sum_{j\in J}\frac{f_j}{Fj}\right)$

Where τ = the highest allowable probability of correctly re-identifying a single record.
f_j = the number of individuals in an equivalence class j in the anonymised dataset.
J= the set of equivalence classes in the anonymised dataset.
$|J|$= the number of unique equivalence classes in the anonymised dataset.
$I(.)$= the indicator function (this returns 1 if the parameter is true and 0 otherwise).
F_j = the number of individuals in an equivalence class j in the matching dataset
n= the total number of records in the anonymised dataset.
N = the total number of records in the matching dataset.

*These metrics are suitable for the situation where the anonymised dataset is a proper subset of a theoretical or actual matching dataset.

- **Calculated example**

This example shows how the calculations were executed on the datasets that were made available through our proposed protocol. Table S4.2 displays a mock de-identified dataset with 25 observations and two independent indirect identifiers: gender_coded and age_group. The variables gender, year_of_birth and age are not expected to be in actual anonymised datasets if gender_coded and age_group are present, but are shown here for completeness.

Appendix 4 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Table S4.2. Mock Anonymised/De-identified dataset (25 observations, 11 unique classes)

ID	Gender	Gender_coded	Year_of_Birth	Age	Age_group
1	Male	1	1933	87	1
2	Male	1	1938	82	1
3	Male	1	1948	72	2
4	Female	2	1949	71	2
5	Female	2	1952	68	3
6	Male	1	1955	65	3
7	Male	1	1956	64	3
8	Female	2	1957	63	3
9	Male	1	1960	60	4
10	Female	2	1963	57	4
11	Male	1	1965	55	4
12	Male	1	1965	55	4
13	Female	2	1965	55	4
14	Female	2	1972	48	5
15	Male	1	1973	47	5
16	Male	1	1974	46	5
17	Female	2	1975	45	5
18	Male	1	1976	44	5
19	Female	2	1980	40	6
20	Male	1	1987	33	6
21	Male	1	1982	38	6
22	Female	2	1983	37	6
23	Male	1	1987	33	6
24	Female	2	1981	39	6
25	Male	1	1984	36	6

Unique class (j)	Gender	Age_group (label)	Number of Observation in each Class (fj)
1	Male	80+	1
2	Male	71-80	2
3	Male	61-70	2
4	Male	51-60	3
5	Male	41-50	3
6	Male	30-40	4
--	Female	80+	0
7	Female	71-80	1
8	Female	61-70	2
9	Female	51-60	2
10	Female	41-50	2
11	Female	30-40	3
TOTAL (n)			25

Prosecutor risk calculation (table S4.3.1) – We obtained three measures of risk under prosecutor risk for the mock anonymised/de-identified dataset presented in table S4.2.

Table S4.3.1

Classes in De-identified dataset and interim calculations for Risk Scores

Prosecutor Risk Scores Calculation*

	Prosecutor risk		tau =	0.33	(Set a priori)	
Unique class (j)	**Gender coded**	**Age_group**	**frequency by class (fj)**	**Individual class risk (1/fj)**	**Is (1/fj)>tau? (Ij)**	**Ij*fj**
1	1	1	1	1.0000	TRUE	1
2	1	2	2	0.5000	TRUE	2
3	1	3	2	0.5000	TRUE	2
4	1	4	3	0.3333	TRUE	3
5	1	5	3	0.3333	TRUE	3
6	1	6	4	0.2500	FALSE	0
--	2	1	0		N/A	N/A
7	2	2	1	1.0000	TRUE	1
8	2	3	2	0.5000	TRUE	2
9	2	4	2	0.5000	TRUE	2
10	2	5	2	0.5000	TRUE	2
11	2	6	3	0.3333	TRUE	3
			25			**21**

R1a	**84%**	The proportion of records that have a re-identification probability higher than 0.33. (R1a=21/25)
R1b	**100%**	Worst case scenario. The class with the highest individual risk represents the whole dataset (R1b=max(1/fj)=1)
R1c	**44%**	Average risk across all of the records in the dataset (Expected value) (R1c=11/25)

*Note that for R1a, the tau (0.33) is set a priory

Appendix 4 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Journalist risk (table S4.3.2) – We assumed that a matching dataset exists (with only 340 observations in order to keep the calculations for this example simple; in reality, matching datasets, when they exist, are much larger) for deterministic matching. We obtained three measures of risk under journalist risk for the mock anonymised/de-identified dataset presented in table S4.2.

Table S4.3.2

Classes in Anonymised/De-identified dataset and interim calculations for Risk Scores

	Journalist risk		tau =	0.09	(Set a priori)			
Unique class (j)	Gender coded	Age_group	frequency by class (fj)	Fj (in matching dataset)	Individual class risk (fj/Fj)	1/Fj	is (1/Fj)>tau? (Ij)	Ij*fj
1	1	1	1	10	0.1000	0.1000	TRUE	1
2	1	2	2	20	0.1000	0.0500	FALSE	0
3	1	3	2	20	0.1000	0.0500	FALSE	0
4	1	4	3	30	0.1000	0.0333	FALSE	0
5	1	5	3	40	0.0750	0.0250	FALSE	0
6	1	6	4	50	0.0800	0.0200	FALSE	0
--	2	1	0	10	0.0000	0.1000	TRUE	0
7	2	2	1	20	0.0500	0.0500	FALSE	0
8	2	3	2	20	0.1000	0.0500	FALSE	0
9	2	4	2	30	0.0667	0.0333	FALSE	0
10	2	5	2	40	0.0500	0.0250	FALSE	0
11	2	6	3	50	0.0600	0.0200	FALSE	0
			25	340	0.8817			1

Journalist Risk Scores Calculations*

R2a	4%	The proportion of records that have a re-identification probability higher than 0.09. (R2a=1/25)
R2b	10%	Worst case scenario. The class with the highest individual risk represents the whole dataset (R2b=max(1/Fj)=0.1)
R2c1	3.5%	R2c1=0.8817/25
R2c2	3.2%	R2c2=11/340
R2c	3.5%	The maximun between the prob of a randomly selected record from the De-identified dataset that is matched against the extra information dataset and the probability of a randomly selected record from the extra information dataset that is matched against the De-identified dataset (R2c=max(3.5,3.2))

*Note that for R2a, a response curve will be generated, as several values of tau will be explored

References

El Emam, K. (2013). Chapter 16 - Measuring the Probability of Re-Identification. Guide to the de-identification of personal health information, CRC Press.

Appendix 5 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendix 5 Data sources requests details and datasets included

Table S5.1 – Data sources requests details and datasets included

id	Repository	Country	Pre selected	Requested	Access Provided		Included Studies
					Controlled	Open	
1	https://datacompass.lshtm.ac.uk [1]	UK	36	7	2	1	NCT02104232[2] NCT02111915[3] ISRCTN36436933[4] ISRCTN74459796[6]
2	https://ctu-app.lshtm.ac.uk/freebird [5]	UK	6	6	--	5	NCT00375258[7] NCT00872469[8] NCT00872469[9] NCT03777488[10] ISRCTN45178534[12] ISRCTN25765518[13]
3	https://datashare.is.ed.ac.uk [11]	UK	31	5	2	3	IST (Registration not required)[14] ISRCTN71907627[15] ISRCTN89489788[16]
4	https://www.clinicalstudydatarequest.com [17]	USA	3058	6	5	--	Registration not required[18] NCT01822899[19] NCT01842607[20] NCT01405053[21] NCT00948766[22]
5	http://datadryad.org [23]	UK	223	5	--	4	ACTRN12616000888460[24] HKCTR-1848[25] No registered[26] NCT04523831[27]
6	http://yoda.yale.edu [28]	USA	410	5	4	--	NCT01715285[29] NCT00903331[30] NCT01004432[31] NCT00211133[32] NCT00058474[34]
7	https://www.projectdatasphere.org [33]	USA	192	9	9	--	NCT00033293[35] NCT00310180[36] NCT00693992[37] NCT00312208[38] NCT00143455[39] NCT00113763[40] NCT00617669[41] NCT00676650[42]
8	https://biolincc.nhlbi.nih.gov/studies [43]	USA	195	6	6	--	NCT00650091[44] NCT00000589[45] NCT00075829[46] NCT01982968[47] NCT01134783[48] NCT00004562[49] NCT00012558[51]
9	https://nda.nih.gov/get/access-data.html [50]	USA	181	6	5	--	Registration not found[52] NCT01927276[53] NCT01944046[54] NCT00005013[55] NCT01198756[57]
10	https://vivli.org/ [56]	USA	3394	8	5	--	NCT01573767[58] NCT01313676[59] NCT00400855[60] NCT01498822[61]
11	https://beta.ukdataservice.ac.uk/datacatalogue/studies [62]	UK	21	5	1	4	ISRCTN11288961[63] ISRCTN90749868[64] NCT01801410[65] Registration not required[66] ISRCTN24081411[67]
12	https://med.data.edu.au/find-data/ [68]	Australia	--	0	--	--	--
13	https://dcri.org/our-approach/data-sharing/soar-data [69]	USA	2	2	--	--	--

Appendix 5 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Table S5.1 – Data sources requests details and datasets selected

id	Repository	Country	Pre selected	Requested	Access Provided		Studies
					Controlled	Open	
14	https://journals.plos.org/plosone/search [70]	UK	5944	6	--	6	NCT02700490[71] TCTR20201005002[72] NCT02185196[73] ACTRN12616000538448[74] ISRCTN 71217488[75] NCT02747524[76]
15	https://www.bmj.com/search/advanced [77]	UK	934	5	--	3	NCT02068885[78] NCT01953549[79] ISRCTN11980540[80] CTRI/2016/09/007240[82]
16	https://dataverse.harvard.edu/ [81]	USA	271	5	--	5	PACTR20190190583260[83] NCT02148952[84] SLCTR/2019/015[85] ANZCTR126160013674[86]
17	https://arlg.org/studies-in-progress/ [87]	USA	--	0	--	--	--
18	https://repository.niddk.nih.gov/studies/ [88]	USA	--	0	--	--	--

Reference List

1. London School of Hygiene & Tropical Medicine. LSHTM Data Compass, https://datacompass.lshtm.ac.uk.
2. Fuhr DC, Weobong B, Lazarus A, et al. Delivering the Thinking Healthy Programme for perinatal depression through peers: an individually randomised controlled trial in India. *The Lancet Psychiatry* 2019; 6: 115-127.
3. Sikander S, Ahmad I, Atif N, et al. Delivering the Thinking Healthy Programme for perinatal depression through volunteer peers: a cluster randomised controlled trial in Pakistan. *The Lancet Psychiatry* 2019; 6: 128-139.
4. Dhalla K, Cousens S, Bowman R, et al. Is beta radiation better than 5 flurouracil as an adjunct for trabeculectomy surgery when combined with cataract surgery? A randomised controlled trial. *PLoS One* 2016; 11: e0161674.
5. Clinical Trials Unit London School of Hygiene & Tropical Medicine. The FreeBIRD Bank of Injury and Emergency Research Data, https://freebird.lshtm.ac.uk/.
6. Collaborators CT. Effect of intravenous corticosteroids on death within 14 days in 10 008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *The Lancet* 2004; 364: 1321-1328.
7. collaborators C-t. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): A randomised, placebo-controlled trial. *the Lancet* 2010; 376: 376: 323-332. DOI: https://doi.org/10.1016/S0140-6736(10)60835-5.
8. Shakur H, Roberts I, Fawole B, et al. Effect of early tranexamic acid administration on mortality, hysterectomy, and other morbidities in women with post-partum haemorrhage (WOMAN): an international, randomised, double-blind, placebo-controlled trial. *The Lancet* 2017; 389: 2105-2116.
9. Shakur-Still H, Roberts I, Fawole B, et al. Effect of tranexamic acid on coagulation and fibrinolysis in women with postpartum haemorrhage (WOMAN-ETAC): a single-centre, randomised, double-blind, placebo-controlled trial. *Wellcome open research* 2018; 3.
10. Grassin-Delyle S, Semeraro M, Lamy E, et al. Pharmacokinetics of tranexamic acid after intravenous, intramuscular, and oral routes: a prospective, randomised, crossover trial in healthy volunteers. *British Journal of Anaesthesia* 2022; 128: 465-472.
11. The University of Edinburgh. Edinburgh DataShare, https://datashare.ed.ac.uk/.
12. Lewis SC, Bhattacharya S, Wu O, et al. Gabapentin for the management of chronic pelvic pain in women (GaPP1): a pilot randomised controlled trial. *PloS one* 2016; 11: e0153037.
13. Group I-C. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *The Lancet* 2012; 379: 2352-2363.
14. Group ISTC. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet* 1997; 349: 1569-1581.
15. Salman RA-S, Dennis M, Sandercock P, et al. Effects of antiplatelet therapy after stroke due to intracerebral haemorrhage (RESTART): a randomised, open-label trial. *The Lancet* 2019; 393: 2613-2623.
16. Mowat C, Arnott I, Cahill A, et al. Mercaptopurine versus placebo to prevent recurrence of Crohn's disease after surgical resection (TOPPIC): a multicentre, double-blind, randomised controlled trial. *The lancet Gastroenterology & hepatology* 2016; 1: 273-282.
17. Clinical Study Data Request (CSDR). Clinical Study Data Request, https://clinicalstudydatarequest.com/ (2020, accessed 26 Oct 2020 2020).
18. Hayden FG, Osterhaus AD, Treanor JJ, et al. Efficacy and safety of the neuraminidase inhibitor zanamivir in the treatment of influenzavirus infections. *New England Journal of Medicine* 1997; 337: 874-880.
19. Singh D, Worsley S, Zhu C-Q, et al. Umeclidinium/vilanterol versus fluticasone propionate/salmeterol in COPD: a randomised trial. *BMC Pulmonary Medicine* 2015; 15: 1-12.
20. Lugogo N, Domingo C, Chanez P, et al. Long-term efficacy and safety of mepolizumab in patients with severe eosinophilic asthma: a multi-center, open-label, phase IIIb study. *Clinical therapeutics* 2016; 38: 2058-2070. e2051.
21. Arzimanoglou A, Ferreira J, Satlin A, et al. Evaluation of long-term safety, tolerability, and behavioral outcomes with adjunctive rufinamide in pediatric patients (≥ 1 to< 4 years old) with Lennox-Gastaut syndrome: final results from randomized study 303. *European Journal of Paediatric Neurology* 2019; 23: 126-135.
22. T Grossberg G, R Farlow M, Meng X, et al. Evaluating high-dose rivastigmine patch in severe Alzheimer's disease: analyses with concomitant memantine usage as a factor. *Current Alzheimer Research* 2015; 12: 53-60.
23. Dryad. Data Dryad, https://datadryad.org/.
24. Darlow B, Stanley J, Dean S, et al. The Fear Reduction Exercised Early (FREE) approach to management of low back pain in general practice: a pragmatic cluster-randomised controlled trial. *PLoS medicine* 2019; 16: e1002897.
25. Zee K-Y, Chan PS, Ho JCS, et al. Adjunctive use of modified Yunu-Jian in the non-surgical treatment of male smokers with chronic periodontitis: a randomized double-blind, placebo-controlled clinical trial. *Chinese Medicine* 2016; 11: 1-13.

Appendix 5 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

26. Christopher PP, Appelbaum PS, Truong D, et al. Reducing therapeutic misconception: A randomized intervention trial in hypothetical clinical trials. *PLoS One* 2017; 12: e0184224.
27. Mahmud R, Rahman MM, Alam I, et al. Ivermectin in combination with doxycycline for treating COVID-19 symptoms: a randomized trial. *Journal of International Medical Research* 2021; 49: 03000605211013550.
28. The Yale University. Yale University Open Data Access (YODA) Project, http://yoda.yale.edu/ (2020, accessed 26 Oct 2020 2020).
29. Fizazi K, Tran N, Fein L, et al. Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. *New England Journal of Medicine* 2017; 377: 352-360.
30. Raghu G, Million-Rousseau R, Morganti A, et al. Macitentan for the treatment of idiopathic pulmonary fibrosis: the randomised controlled MUSIC trial. *European Respiratory Journal* 2013; 42: 1622-1632.
31. Huffstutter JE, Kafka S, Brent LH, et al. Clinical response to golimumab in rheumatoid arthritis patients who were receiving etanercept or adalimumab: results of a multicenter active treatment study. *Current Medical Research and Opinion* 2017; 33: 657-666.
32. Leyland-Jones B, Semiglazov V, Pawlicki M, et al. Maintaining normal hemoglobin levels with epoetin alfa in mainly nonanemic patients with metastatic breast cancer receiving first-line chemotherapy: a survival study. *Journal of Clinical Oncology* 2005; 23: 5960-5972.
33. CEO Roundtable on Cancer Inc. Project Data Sphere, https://www.projectdatasphere.org/.
34. O'Connell MJ, Colangelo LH, Beart RW, et al. Capecitabine and oxaliplatin in the preoperative multimodality treatment of rectal cancer: surgical end points from National Surgical Adjuvant Breast and Bowel Project trial R-04. *Journal of clinical oncology* 2014; 32: 1927.
35. de Alarcon PA, Matthay KK, London WB, et al. Intravenous immunoglobulin with prednisone and risk-adapted chemotherapy for children with opsoclonus myoclonus ataxia syndrome associated with neuroblastoma (ANBL00P3): a randomised, open-label, phase 3 trial. *The Lancet Child & Adolescent Health* 2018; 2: 25-34.
36. Sparano JA, Gray RJ, Makower DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine* 2018; 379: 111-121.
37. Baggstrom MQ, Socinski MA, Wang XF, et al. Maintenance sunitinib following initial platinum-based combination chemotherapy in advanced-stage IIIB/IV non–small cell lung cancer: a randomized, double-blind, placebo-controlled phase III study—CALGB 30607 (Alliance). *Journal of Thoracic Oncology* 2017; 12: 843-849.
38. Eiermann W, Pienkowski T, Crown J, et al. Phase III study of doxorubicin/cyclophosphamide with concomitant versus sequential docetaxel as adjuvant treatment in patients with human epidermal growth factor receptor 2-normal, node-positive breast cancer: BCIRG-005 trial. *J Clin Oncol* 2011; 29: 3877-3884.
39. Pfizer. *Online report for Open Label, Randomised Multicentre Phase III Study Of Irinotecan Hydrochloride (Campto (Registered)) And Cisplatin Versus Etoposide And Cisplatin In Chemotherapy Naive Patients With Extensive Disease - Small Cell Lung Cancer*. 2010.
40. Poulin-Costello M, Azoulay L, Van Cutsem E, et al. An analysis of the treatment effect of panitumumab on overall survival from a phase 3, randomized, controlled, multicenter trial (20020408) in patients with chemotherapy refractory metastatic colorectal cancer. *Targeted oncology* 2013; 8: 127-136.
41. Fizazi K, Higano CS, Nelson JB, et al. Phase III, randomized, placebo-controlled study of docetaxel in combination with zibotentan in patients with metastatic castration-resistant prostate cancer. *Journal of Clinical Oncology* 2013; 31: 1740-1747.
42. Michaelson MD, Oudard S, Ou Y-C, et al. Randomized, placebo-controlled, phase III trial of sunitinib plus prednisone versus prednisone alone in progressive, metastatic, castration-resistant prostate cancer. *J Clin Oncol* 2014; 32: 76-82.
43. The National Heart LaBIN. BioLINCC, https://biolincc.nhlbi.nih.gov/.
44. Network IPFCR. Randomized trial of acetylcysteine in idiopathic pulmonary fibrosis. *New England Journal of Medicine* 2014; 370: 2093-2101.
45. Group TtRAtPS. Leukocyte reduction and ultraviolet B irradiation of platelets to prevent alloimmunization and refractoriness to platelet transfusions. *New England Journal of Medicine* 1997; 337: 1861-1870.
46. Krishnan A, Pasquini MC, Logan B, et al. Autologous haemopoietic stem-cell transplantation followed by allogeneic or autologous haemopoietic stem-cell transplantation in patients with multiple myeloma (BMT CTN 0102): a phase 3 biological assignment trial. *The lancet oncology* 2011; 12: 1195-1203.
47. Raghu G, Pellegrini CA, Yow E, et al. Laparoscopic anti-reflux surgery for the treatment of idiopathic pulmonary fibrosis (WRAP-IPF): a multicentre, randomised, controlled phase 2 trial. *The Lancet Respiratory Medicine* 2018; 6: 707-714.
48. Lytle LA, Laska MN, Linde JA, et al. Weight-gain reduction among 2-year college students: the CHOICES RCT. *American Journal of Preventive Medicine* 2017; 52: 183-191.
49. Hochman JS, Lamas GA, Buller CE, et al. Coronary intervention for persistent occlusion after myocardial infarction. *New England Journal of Medicine* 2006; 355: 2395-2407.
50. The National Institute of Mental Health. The NIMH Data Archive (NDA), https://nda.nih.gov/.
51. Sachs GS, Nierenberg AA, Calabrese JR, et al. Effectiveness of adjunctive antidepressant treatment for bipolar depression. *New England Journal of Medicine* 2007; 356: 1711-1722.
52. Kerwin ML. Using SMART Treatment Design to Evaluate Applied Behavior Analysis Interventions on Communication in Preschool Children with Autism.
53. Kelly DL, Demyanovich HK, Rodriguez KM, et al. Randomized controlled trial of a gluten-free diet in patients with schizophrenia positive for antigliadin antibodies (AGA IgG): a pilot feasibility study. *Journal of Psychiatry and Neuroscience* 2019; 44: 269-276.
54. Sikich L, Kolevzon A, King BH, et al. Intranasal oxytocin in children and adolescents with autism spectrum disorder. *New England Journal of Medicine* 2021; 385: 1462-1473.
55. Group HDTS and Group HDTS. Effect of Hypericum perforatum (St John's wort) in major depressive disorder: a randomized controlled trial. *Jama* 2002; 287: 1807-1814.
56. Vivli Center for Global Clinical Research Data. Vivli, a global data-sharing and analytics platform. , https://vivli.org/ (2020, accessed 30 Oct 2020 2020).
57. Langley JM, Carmona Martinez A, Chatterjee A, et al. Immunogenicity and safety of an inactivated quadrivalent influenza vaccine candidate: a phase III randomized controlled trial in children. *The Journal of infectious diseases* 2013; 208: 544-553.
58. Oliver AJ, Covar RA, Goldfrad CH, et al. Randomised trial of once-daily vilanterol in children with asthma on inhaled corticosteroid therapy. *Respiratory Research* 2016; 17: 1-11.
59. Calverley PM, Anderson JA, Brook RD, et al. Fluticasone furoate, vilanterol, and lung function decline in patients with moderate chronic obstructive pulmonary disease and heightened cardiovascular risk. *American Journal of Respiratory and Critical Care Medicine* 2018; 197: 47-55.
60. GlaxoSmithKline group of companies. *A Randomised, Double-blind, Placebo-controlled, Incomplete Block, 4-period Crossover, Study to Investigate the Effects of 5-day Repeat Inhaled Doses of Fluticasone Propionate (BID, 50-2000 mcg) on Airway Responsiveness to Adenosine 5-monophosphate (AMP) Challenge When Delivered After the Last Dose in Mild Asthmatic Subjects. (GSK 04_SIG103337). Clinical Summary Report* 20 October 2006 2006.
61. Kim JH, Lee SK, Loesch C, et al. Comparison of levetiracetam and oxcarbazepine monotherapy among Korean patients with newly diagnosed focal epilepsy: A long-term, randomized, open-label trial. *Epilepsia* 2017; 58: e70-e74.
62. UK Data Service - University of Essex. UK Data Service: data Catalogue, https://beta.ukdataservice.ac.uk/datacatalogue/studies.
63. Csipke E, Shafayat A, Sprange K, et al. Promoting independence in dementia (PRIDE): A feasibility randomized controlled trial. *Clinical Interventions in Aging* 2021: 363-378.

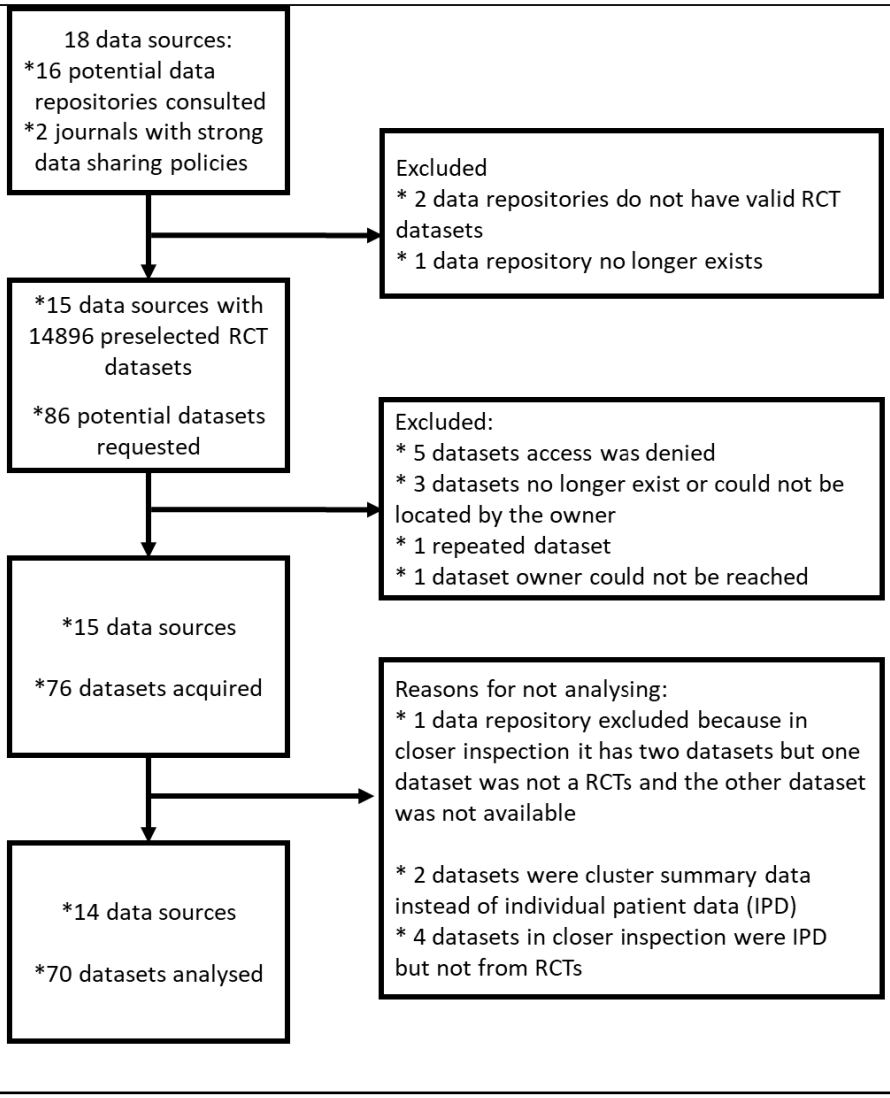
Appendix 5 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

64. Yiend J, Lam CL, Schmidt N, et al. Cognitive bias modification for paranoia (CBM-pa): a randomised controlled feasibility study in patients with distressing paranoid beliefs. *Psychological medicine* 2023; 53: 4614-4626.
65. Bracken H, Mundle S, Faragher B, et al. Induction of labour in pre-eclamptic women: a randomised trial comparing the Foley balloon catheter with oral misoprostol. *BMC pregnancy and childbirth* 2014; 14: 1-5.
66. McEwan K, Richardson M, Sheffield D, et al. A smartphone app for improving mental health through connecting with urban nature. *International journal of environmental research and public health* 2019; 16: 3373.
67. Murphy AW, Cupples M, Smith S, et al. Effect of tailored practice and patient care plans on secondary prevention of heart disease in general practice: cluster randomised controlled trial. *Bmj* 2009; 339.
68. Intersect Australia Limited - Queensland Cyber Infrastructure Foundation Ltd. Australian National Medical Research Data Storage Facility, https://med.data.edu.au/find-data/.
69. Institute DCR. SOAR DATA™, https://dcri.org/our-approach/data-sharing/soar-data.
70. PLOS is a nonprofit 501(c)(3) corporation. PLOS ONE: An inclusive journal community working together to advance science by making all rigorous research accessible without barriers, https://journals.plos.org/plosone/search.
71. Anjara SG, Bonetto C, Ganguli P, et al. Can General Practitioners manage mental disorders in primary care? A partially randomised, pragmatic, cluster trial. *PLoS One* 2019; 14: e0224724.
72. Vijitpavan A, Kittikunakorn N and Komonhirun R. Comparison between intrathecal morphine and intravenous patient control analgesia for pain control after video-assisted thoracoscopic surgery: A pilot randomized controlled study. *Plos one* 2022; 17: e0266324.
73. Chowdhury F, Shahid ASMSB, Tabassum M, et al. Vitamin D supplementation among Bangladeshi children under-five years of age hospitalised for severe pneumonia: A randomised placebo controlled trial. *Plos one* 2021; 16: e0246460.
74. Weinberg L, Ianno D, Churilov L, et al. Restrictive intraoperative fluid optimisation algorithm improves outcomes in patients undergoing pancreaticoduodenectomy: a prospective multicentre randomized controlled trial. *PLoS One* 2017; 12: e0183313.
75. Choi W, Kim JC, Kim WS, et al. Clinical effect of antioxidant glasses containing extracts of medicinal plants in patients with dry eye disease: a multi-center, prospective, randomized, double-blind, placebo-controlled trial. *PLoS One* 2015; 10: e0139761.
76. Iannotti L, Dulience SJ-L, Joseph S, et al. Fortified snack reduced anemia in rural school-aged children of Haiti: a cluster-randomized, controlled trial. *PloS one* 2016; 11: e0168121.
77. BMJ Publishing Group Ltd. BMJ is a global healthcare knowledge provider with a vision for a healthier world. We share knowledge and expertise to improve healthcare outcomes., https://www.bmj.com/search/advanced.
78. Ebbeling CB, Feldman HA, Klein GL, et al. Effects of a low carbohydrate diet on energy expenditure during weight loss maintenance: randomized trial. *bmj* 2018; 363.
79. Nave AH, Rackoll T, Grittner U, et al. Physical Fitness Training in Patients with Subacute Stroke (PHYS-STROKE): multicentre, randomised controlled, endpoint blinded trial. *Bmj* 2019; 366.
80. Costa ML, Achten J, Ooms A, et al. Surgical fixation with K-wires versus casting in adults with fracture of distal radius: DRAFFT2 multicentre randomised clinical trial. *bmj* 2022; 376.
81. Harvard University. Harvard Dataverse Repository. Deposit and share your data. Get academic credit. Harvard Dataverse is a repository for research data. Deposit data and code here., https://dataverse.harvard.edu/.
82. Gehani M, Kapur S, Madhuri SD, et al. Effectiveness of antenatal screening of asymptomatic bacteriuria in reduction of prematurity and low birth weight: Evaluating a point-of-care rapid test in a pragmatic randomized controlled study. *EClinicalMedicine* 2021; 33.
83. Elson L, Randu K, Feldmeier H, et al. Efficacy of a mixture of neem seed oil (Azadirachta indica) and coconut oil (Cocos nucifera) for topical treatment of tungiasis. A randomized controlled, proof-of-principle study. *PLoS Neglected Tropical Diseases* 2019; 13: e0007822.
84. Semrau KE, Hirschhorn LR, Marx Delaney M, et al. Outcomes of a coaching-based WHO safe childbirth checklist program in India. *New England Journal of Medicine* 2017; 377: 2313-2324.
85. Jayawardane M, Piyadigama I and Chandradeva U. Will a preoperative theatre visit reduce anxiety? A randomised controlled trial. *Journal of Obstetrics and Gynaecology* 2022; 42: 1498-1503.
86. Stitely ML, Harlow K and MacKenzie E. Oral riboflavin to assess ureteral patency during cystoscopy: a randomized clinical trial. *Obstetrics & Gynecology* 2019; 133: 301-307.
87. Antibacterial Resistance Leadership Group (ARLG) ARLG studies, https://arlg.org/summary-of-results/.
88. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). NIDDK Central Repository, https://repository.niddk.nih.gov/studies/dpp/.

Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendix 6 Additional figures and tables

Figure S6.1.1 Flow chart for the selection and analysis for clinical trials' datasets

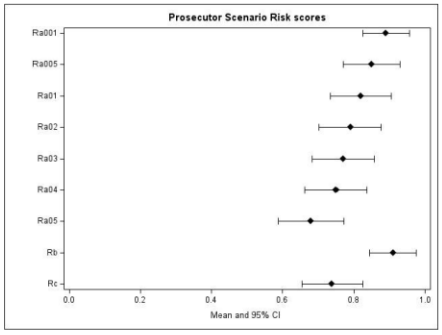
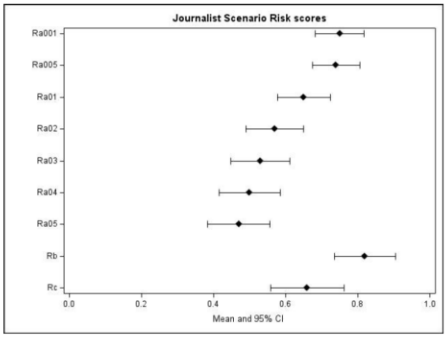
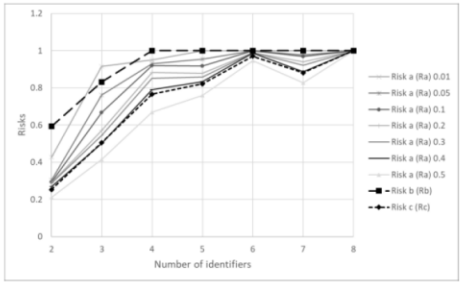
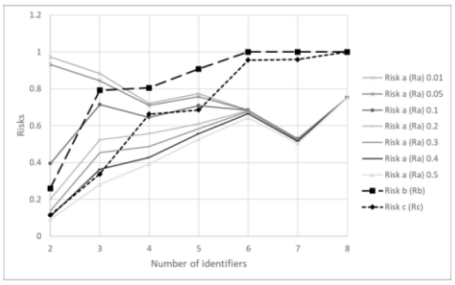
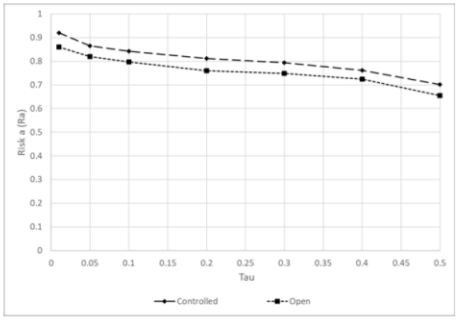
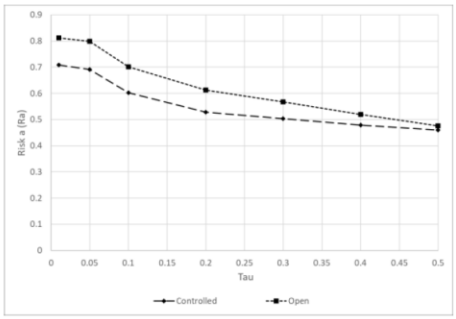


Where:

RCT	Randomised controlled trial
IPD	Individual patient data

Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Figure. S6.1.2 Mean re-identification Risks scores

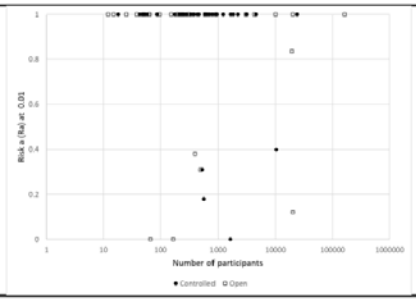
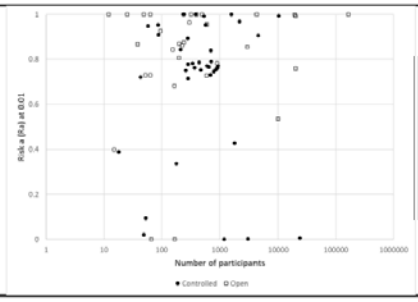
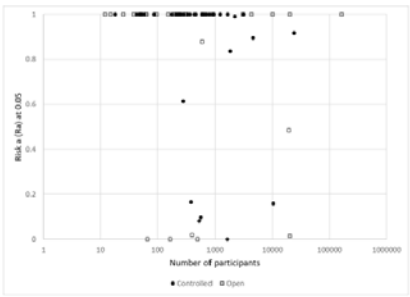
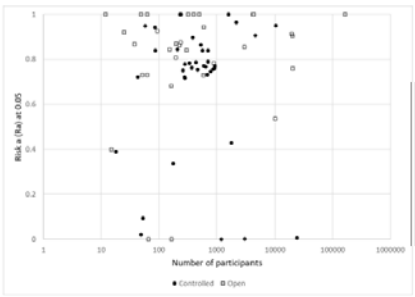
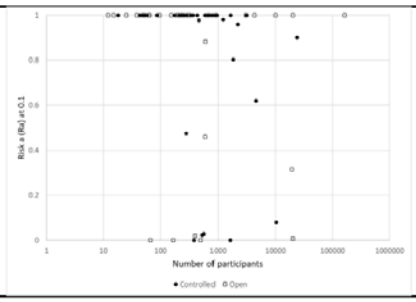
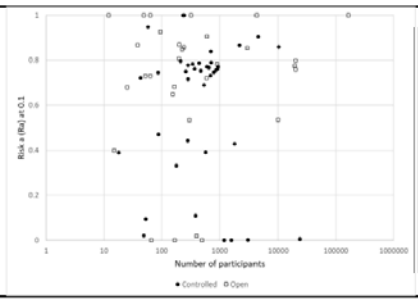
	Prosecutor scenario	Journalist scenario
Mean with 95% CI		
By number of identifiers in clinical trial datasets	 Note: automatic calculations of risk scores are not included in this graphs	 Note: automatic calculations of risk scores are not included in this graphs
Risks a (Ra) vs levels of threshold by scenario		

Where: Risk a (Ra): the proportions of participants in strata above a predetermined risk threshold, Risk b (Rb): the stratum with the smallest membership in the anonymised dataset, and Risk c (Rc): the average risk score across the whole strata of the anonymised dataset, using all indirect identifiers, calculated with the formulas in chapter 16 from "Guide to the de-identification of personal health information" by Khaled El Emam (2013)

Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

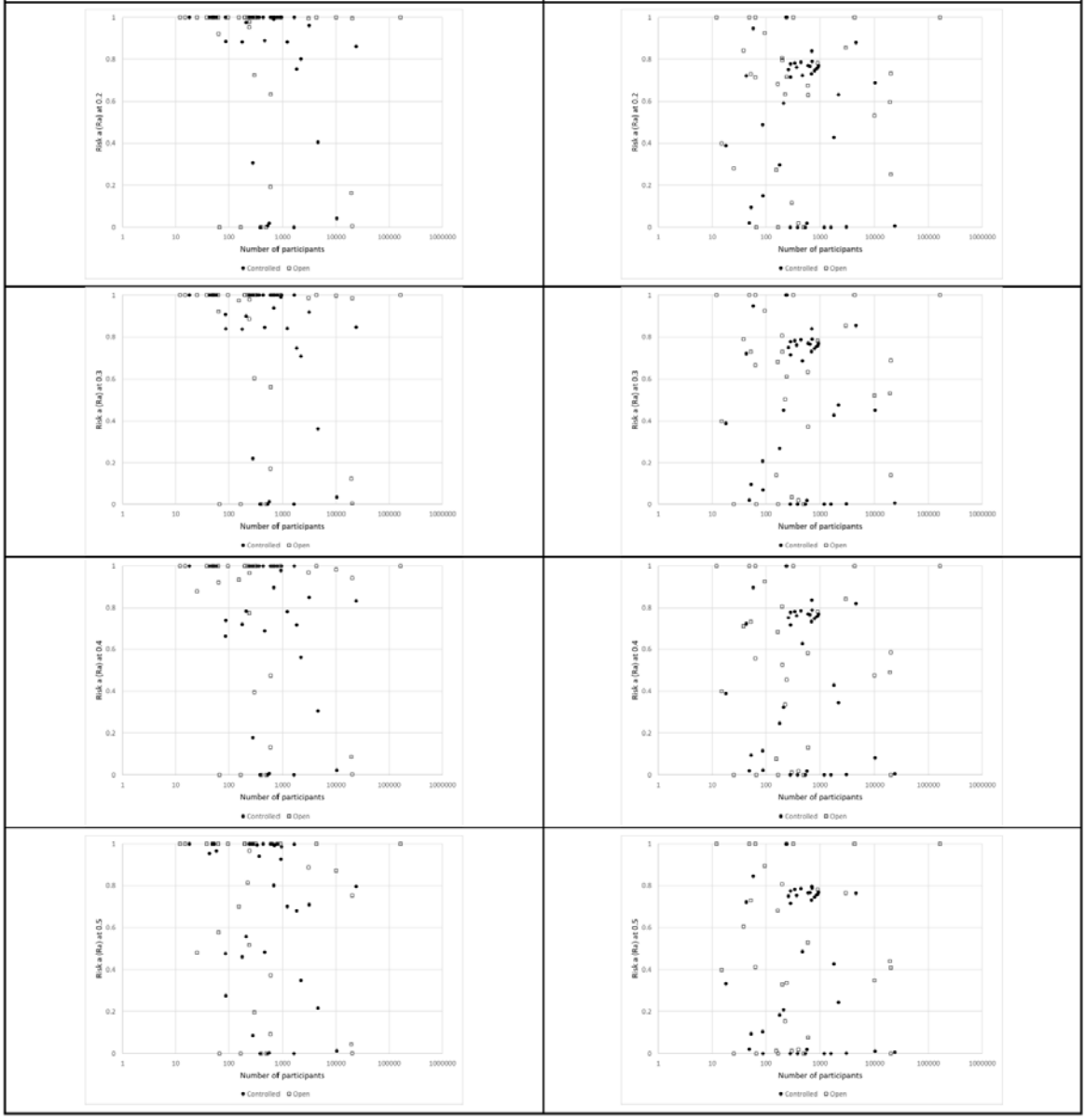
Index of Pre-planned Plots for Re-identification Risk Scores in clinical trials' datasets	
Figure S6.2.1	Scatterplots of Ra vs anonymised clinical trial datasets' sample size
Figure S6.2.2	Box-and-whisker plots of Ra vs number of indirect identifiers
Figure S6.2.3	Scatterplot of Rb vs anonymised clinical trial datasets' sample size
Figure S6.2.4	Box-and-whisker plots of Rb vs number of indirect identifiers
Figure S6.2.5	Scatterplot of Rc vs anonymised clinical trial datasets' sample size
Figure S6.2.6	Box-and-whisker plots of Rc vs number of indirect identifiers
Figure S6.2.7	Scatterplots of Ra vs Rb
Figure S6.2.8	Scatterplots of Ra vs Rc
Figure S6.2.9	Scatterplots of Rb vs Rc
Where:	Risk a (Ra): the proportions of participants in strata above a predetermined risk threshold, Risk b (Rb): the stratum with the smallest membership in the anonymised dataset, and Risk c (Rc): the average risk score across the whole strata of the anonymised dataset, using all indirect identifiers, calculated with the formulas in chapter 16 from "Guide to the de-identification of personal health information" by Khaled El Emam (2013)

Figure S6.2.1	Scatterplots of Ra vs anonymised clinical trial datasets' sample size

Prosecutor scenario	Journalist scenario
	
	
	

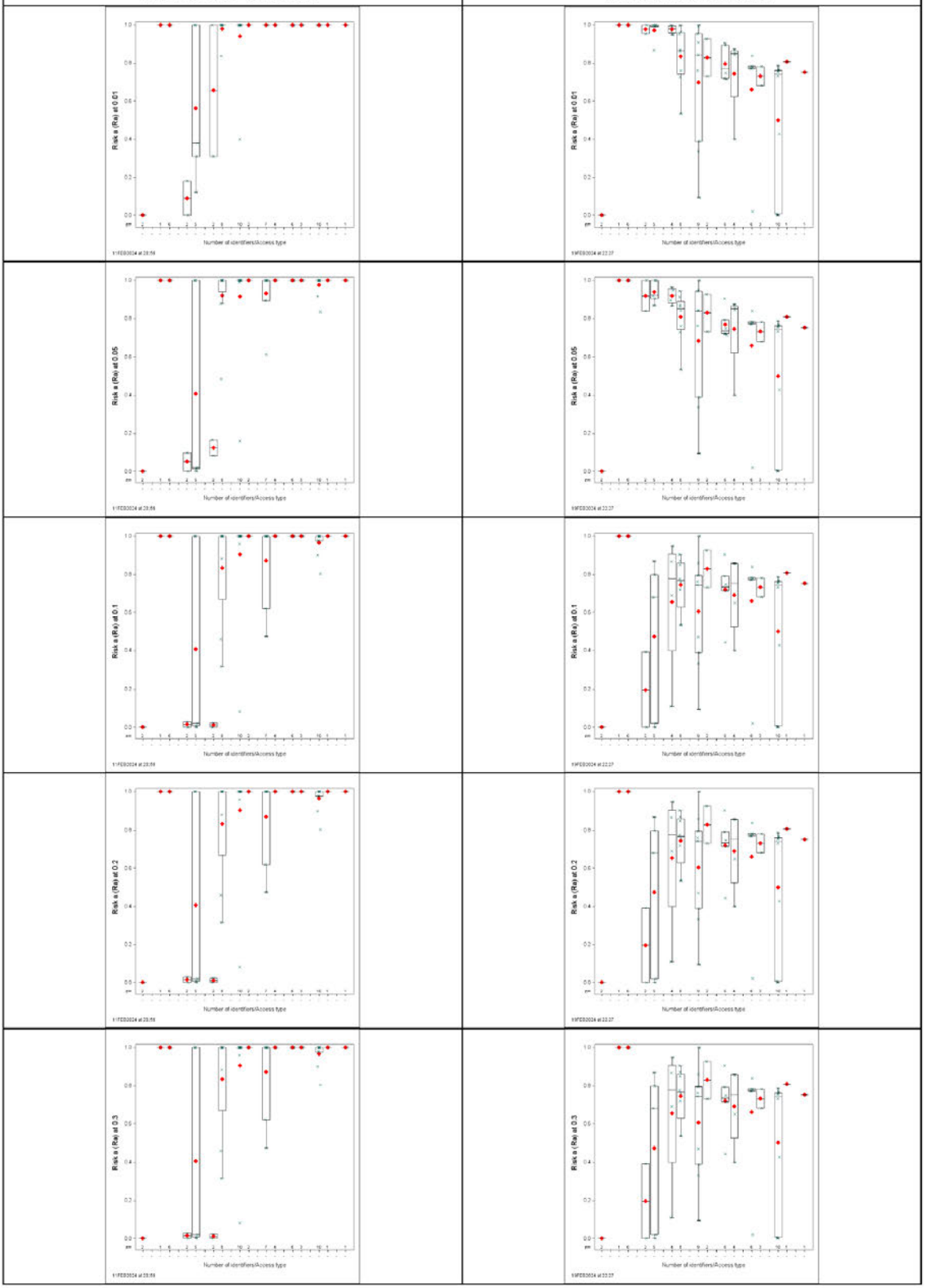
Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Figure S6.2.1	Scatterplots of Ra vs anonymised clinical trial datasets' sample size
Prosecutor scenario	Journalist scenario



Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Figure S6.2.2	Box-and-whisker plots of Ra vs number of indirect identifiers
Prosecutor scenario	Journalist scenario



Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Figure S6.2.2	Box-and-whisker plots of Ra vs number of indirect identifiers

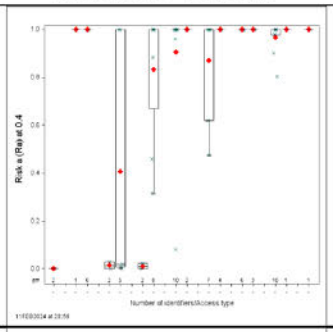
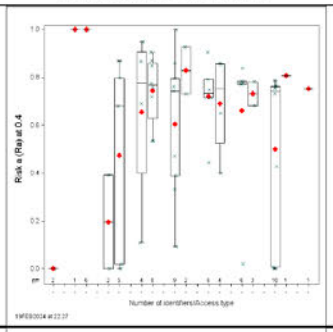
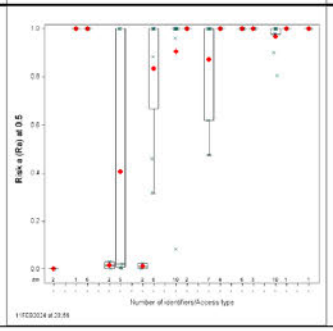
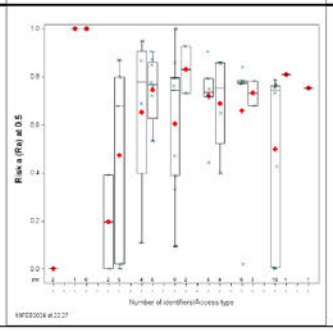
Prosecutor scenario	Journalist scenario
	
	

Figure S6.2.3	Scatterplots of Rb vs anonymised clinical trial datasets' sample size

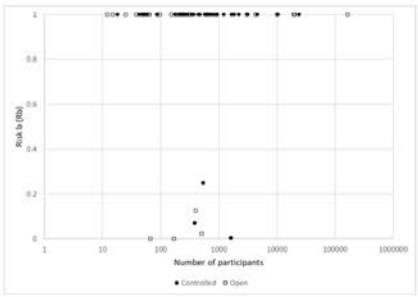
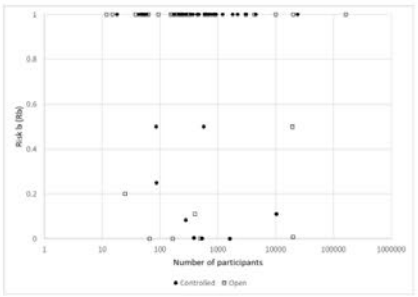
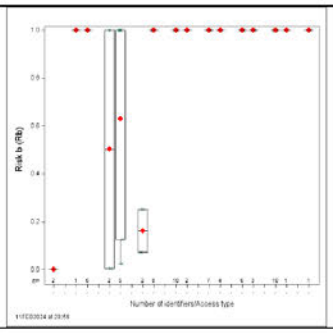
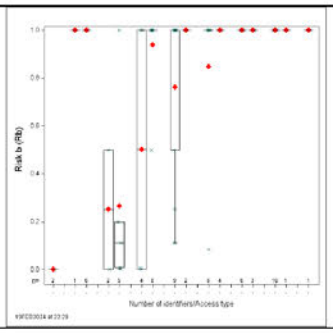
Prosecutor scenario	Journalist scenario
	

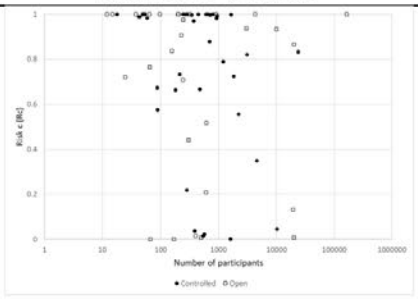
Figure S6.2.4	Box-and-whisker plots of Rb vs number of indirect identifiers

Prosecutor scenario	Journalist scenario
	

Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Figure S6.2.5	Scatterplots of Rc vs anonymised clinical trial datasets' sample size

Prosecutor scenario



Journalist scenario

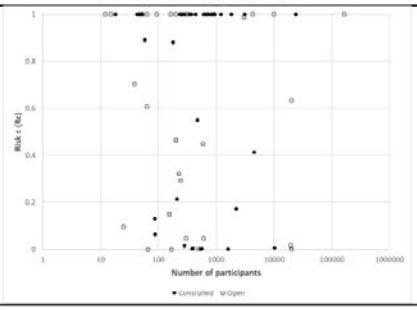
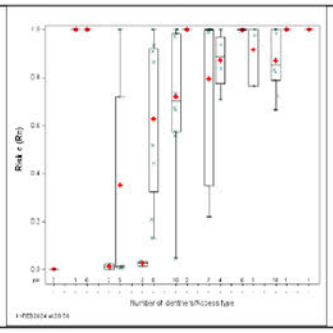


Figure S6.2.6	Box-and-whisker plots of Rc vs number of indirect identifiers

Prosecutor scenario



Journalist scenario

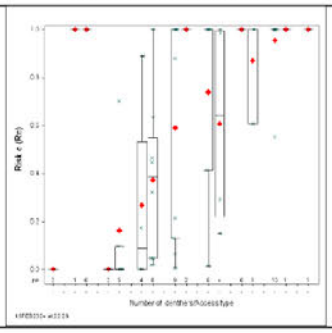
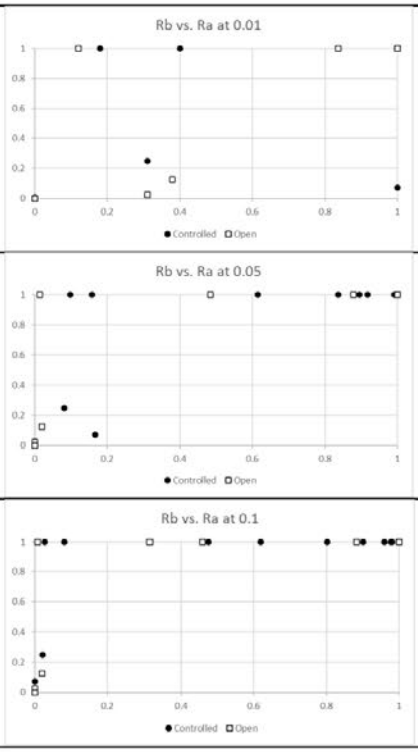
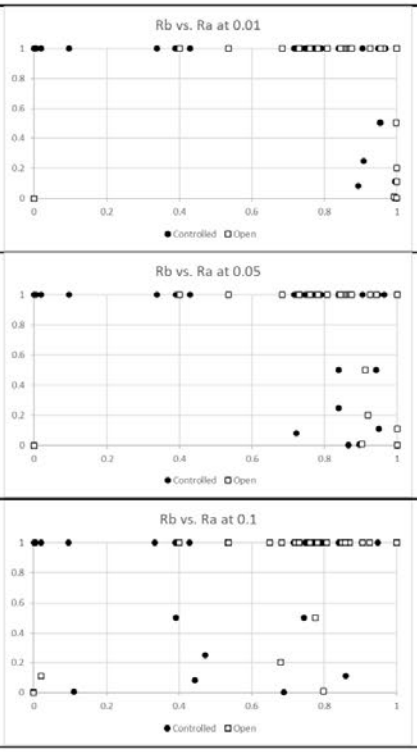


Figure S6.2.7	Scatterplots of Rb vs Ra

Prosecutor scenario



Journalist scenario



Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Figure S6.2.7	Scatterplots of Rb vs Ra

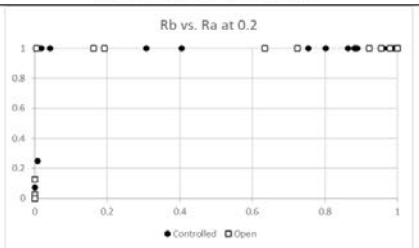
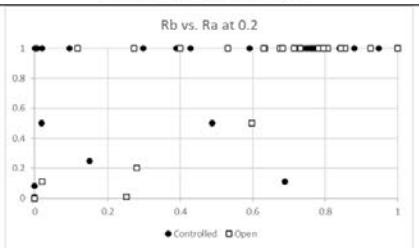
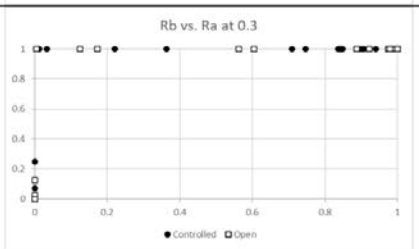
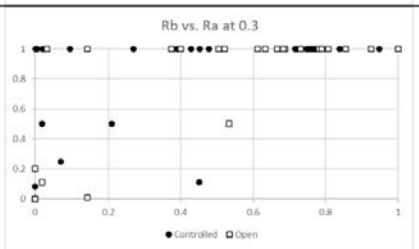
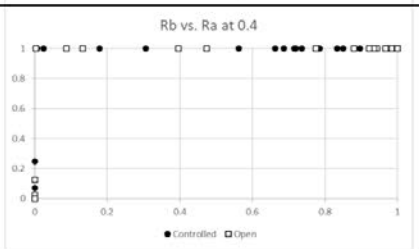
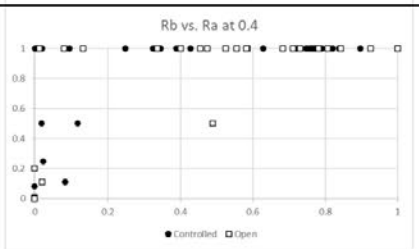
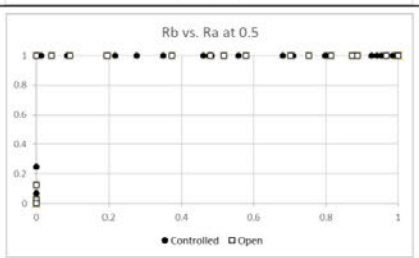
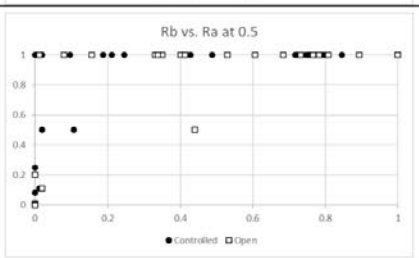
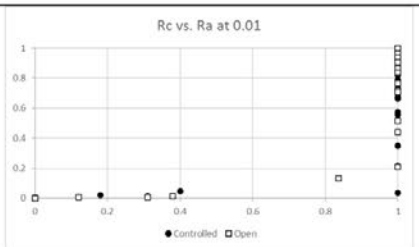
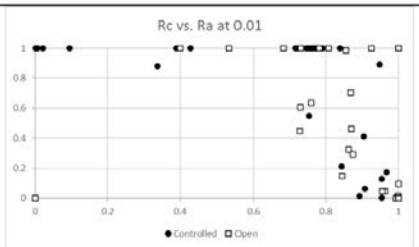
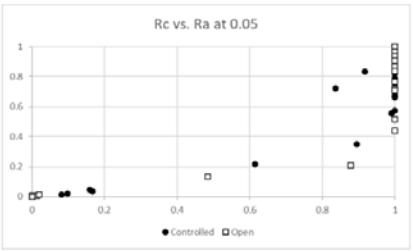
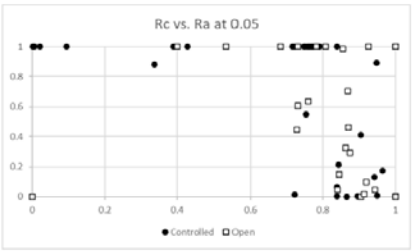
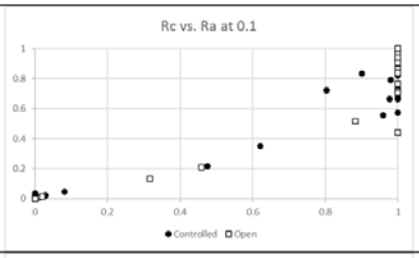
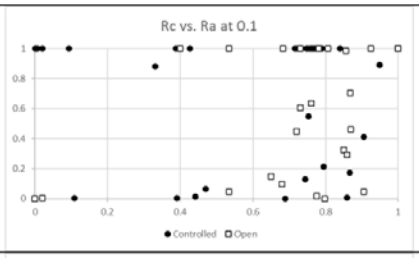
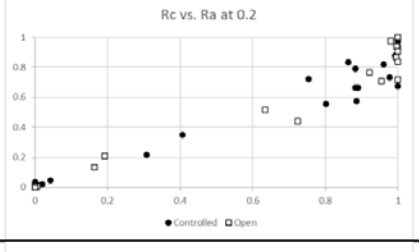
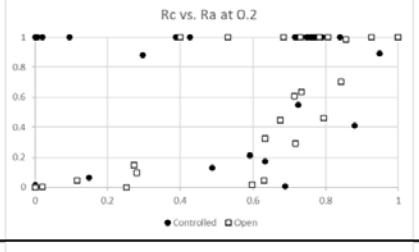
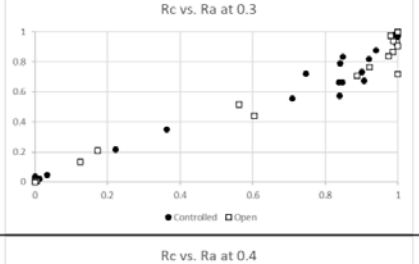
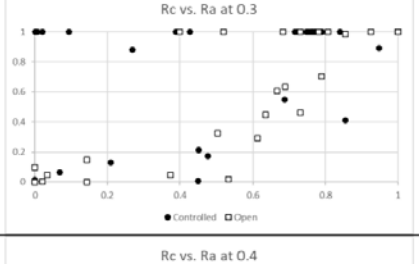
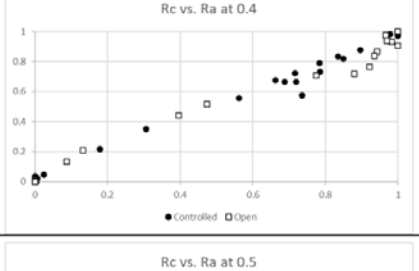
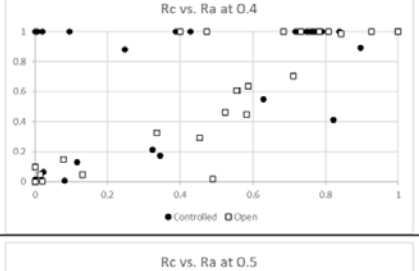
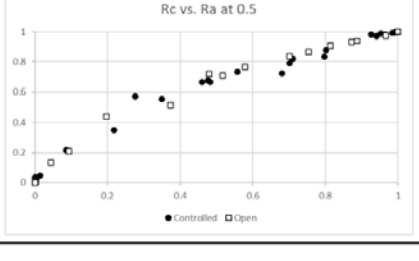
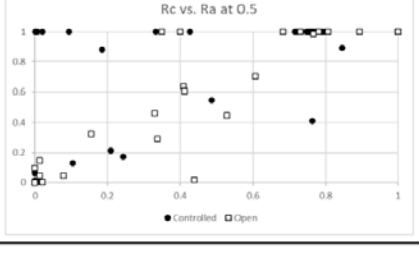
Prosecutor scenario	Journalist scenario
	
	
	
	

Figure S6.2.8	Scatterplots of Rc vs Ra

Prosecutor scenario	Journalist scenario
	

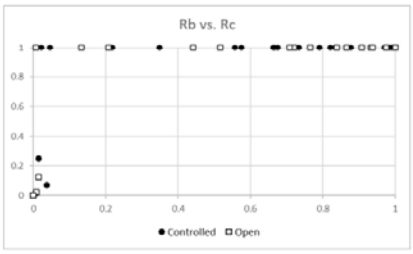
Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

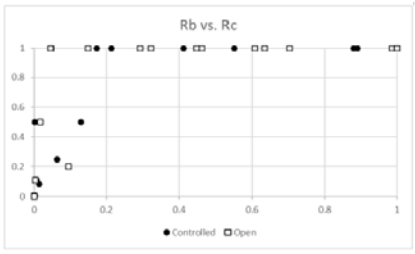
Figure S6.2.8	Scatterplots of Rc vs Ra

Prosecutor scenario	Journalist scenario
	
	
	
	
	
	

Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Figure S6.2.9	Scatterplot of Rb vs Rc
Prosecutor scenario	Journalist scenario





Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Table S6.3 – **Toppic[1] trial re-identification risk scores under prosecutor scenario.**

	Indirect Identifiers					Re-identification risk scores								
Categorisation	Age (scale / participants)	Sex (scale / participants)	Height (scale / participants)	Weight (scale / participants)	Unique levels	Ra 0.01	Ra 0.05	Ra 0.1	Ra 0.2	Ra 0.3	Ra 0.4	Ra 0.5	Rb	Rc
Original Anonymised dataset	235 levels	2 levels: F 146, M 94	33 levels	159 levels	240	1	1	1	1	1	1	1	1	1
Fine - continuous indirect identifiers (age, height and weight) categorised in bands of 10 units	8 levels: <18 4; 18-24 36; 25-34 64; 35-44 45; 45-54 53; 55-64 28; 65-74 4; Missing 6	2 levels: F 146, M 94	6 levels: <150 3; 150-159 40; 160-169 89; 170-179 76; 180-189 29; 190-199 3	9 levels: <50 7; 50-59 51; 60-69 62; 70-79 52; 80-89 34; 90-99 14; 100-109 6; 110>= 3; Missing 11	123	1	1	1	0.82	0.75	0.49	0.27	1	0.51
Medium - continuous indirect identifiers (age, height and weight) categorised in bands of 20 units	5 levels: <25 40; 25-44 109; 45-64 81; 65-84 4; Missing 6	2 levels: F 146, M 94	4 levels: <150 3; 150-169 129; 170-189 105; 190>= 3	5 levels: <60 58; 60-79 114; 80-99 48; 100>= 9; Missing 11	52	1	0.73	0.50	0.29	0.24	0.15	0.07	1	0.22
Coarse – as medium but strata with counts less than 5 collapsed with their most adjacent stratum	4 levels: <25 40; 25-44 109; 45>= 85; Missing 6	2 levels: F 146, M 94	2 levels: <170 132; 170>= 108	5 levels: <60 58; 60-79 114; 80-99 48; 100>= 9; Missing 11	44	1	0.70	0.46	0.26	0.18	0.13	0.05	1	0.18
Coarse with age removed	Removed	2 levels: F 146, M 94	2 levels: <170 132; 170>= 108	5 levels: <60 58; 60-79 114; 80-99 48; 100>= 9; Missing 11	17	1	0.31	0.20	0.04	0.04	0.02	0.01	1	0.07
Coarse with weight removed	4 levels: <25 40; 25-44 109; 45>= 85; Missing 6	2 levels: F 146, M 94	2 levels: <170 132; 170>= 108	Removed	15	1	0.30	0.13	0.08	0.03	0.01	0.004	1	0.06

Notes: 1 Mowat C, Arnott I, Cahill A, et al. Mercaptopurine versus placebo to prevent recurrence of Crohn's disease after surgical resection (TOPPIC): a multicentre, double-blind, randomised controlled trial. The lancet Gastroenterology & hepatology 2016; 1: 273-282. (This trial has 240 participants)

Where: Risk a (Ra): the proportions of participants in strata above a predetermined risk threshold, Risk b (Rb): the stratum with the smallest membership in the anonymised dataset, and Risk c (Rc): the average risk score across the whole strata of the anonymised dataset, using all indirect identifiers, calculated with the formulas in chapter 16 from "Guide to the de-identification of personal health information" by Khaled El Emam (2013)

Appendix 6 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Table S6.4 – **Re-identification risk scores under prosecutor scenario for two datasets with three indirect identifiers.**

	Indirect identifiers				Number of participants	Unique levels	Re-identification risk scores								
ID	Age	Sex	Country	Ethnicity			Ra 0.01	Ra 0.05	Ra 0.1	Ra 0.2	Ra 0.3	Ra 0.4	Ra 0.5	Rb	Rc
Dataset 1[1] (RESTART)	2 levels	2 levels	--	2 levels	537	8	0.31	0.08	0.02	0.007	0	0	0	0.25	0.015
Dataset 2[2] (IST)	82 levels	2 levels	32 levels	--	19435	2570	0.83	0.48	0.31	0.16	0.12	0.09	0.04	1	0.13

Notes:
1 Salman, R.A.-S., et al., Effects of antiplatelet therapy after stroke due to intracerebral haemorrhage (RESTART): a randomised, open-label trial. The Lancet, 2019. 393(10191): p. 2613-2623
2 Group, I.S.T.C., The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. The Lancet, 1997. 349(9065): p. 1569-1581

Where: Risk a (Ra): the proportions of participants in strata above a predetermined risk threshold, Risk b (Rb): the stratum with the smallest membership in the anonymised dataset, and Risk c (Rc): the average risk score across the whole strata of the anonymised dataset, using all indirect identifiers, calculated with the formulas in chapter 16 from "Guide to the de-identification of personal health information" by Khaled El Emam (2013)

Appendix 7 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

Appendix 7 Ethics/data protection application

USHER RESEARCH ETHICS GROUP 'TRIAGE': VERSION OF FORM TO USE FOR ETHICS APPLICATION

It is Usher Institute (UI) policy that ethics oversight is required for ***all Usher research projects*** carried out by staff or students of Usher, that utilise individual data from living persons. Many studies will obtain ethics approval from NHS REC, while GCRF studies have a special process co-ordinated at College MVM level. In respect of *all other Usher research*, an **ethics oversight service** is offered by the Usher Research Ethics Group (UREG) offers. However this service is **open only to staff and PG students of the Usher Institute**, not to UI affiliates or their PG students.

The Usher Research Ethics Group (UREG) is part of the **School of Health in Social Science** (SHiSS) **Research Ethics Committee**, within the governance framework of the College of Arts, Humanities and Social Sciences (CAHSS). Since July 2019 UREG has had an integrated **Usher Registration and Ethics form** which combines the old UREG ethics forms with the 'mini' DPIA [created post-GDPR (May 2018). This is close to what will pertain in the forthcoming University-wide ethics oversight software, which is being commissioned by UoE.

However, in its current 'flat document' multi-purpose version, the single UREG ethics application form can seem obtuse and complex to use... We have therefore devised this initial **Triage Flow Chart** stage, to help you identify up front *how much* of the form you need to complete, and after using this flow chart, to enable you to select the version of the form most suited, and specific, to your intended research. This should make completion of your ethics application much simpler for you.

This has been possible because, under UREG process, *how much* information/reflection/explanation needs to be provided in the ethics application form, is 'proportionate', and will depend on the *nature* of your research (primary/secondary) and your *data*. Formal 'ethics' review will be required only for:

- Primary data collection studies and
- A subset of secondary data analysis (SDA) studies (this is a consequence of GDPR, and encompasses only those that do not utilise 'fully anonymous' data).

  [However, SDA studies that do require UREG oversight will require *only a limited degree of* ethics oversight, mainly focussing on data protection issues, and this may be minimal if a DPIA has already been completed.]

However, a consequence of this simplification of forms is that you may no longer combine, *on one form*, both primary research (new data collection) ***and* secondary data analysis**. If this is a problem for you, please get in touch with UREG at usher.ethics@ed.ac.uk

WHAT TO DO

All research projects collecting/utilising individual data, are required to obtain ethics approval ***prior to*** *any commencement of the research project*. If you are not going through NHS REC ethics review, please use the flow chart overleaf to identify whether you need to submit an ethics application to Usher Ethics Committee, and if so, the version of the UREG ethics application form you need to use.

Download the identified form and follow its instructions for completion. Note that you are *very likely* to need to enquire to ACCORD about UoE sponsorship, *before* submitting your ethics application to UREG.

Ethics resources are available in Usher ethics shared area.[1] In particular, please see the doc **UREG Guidance manual to ethics oversight**, and the doc **Useful UoE paths and URLs for ethics**, which points to myriad further resources available within subfolder ***Other useful ethics docs***.

Ultimately, your application submission should include *both* this triage sheet, *and* your completed application form (plus any other documents that might be needed). Submission please, and any **queries,** to usher.ethics@ed.ac.uk.

Note to PG/PGR student applicants and their supervisors:

The flow chart use needs to be completed *under the oversight* of the PG/PGR student's UoE supervisor(s). ***Responsibility for the validity and accuracy of the triage stage lies with the academic supervisor(s).***

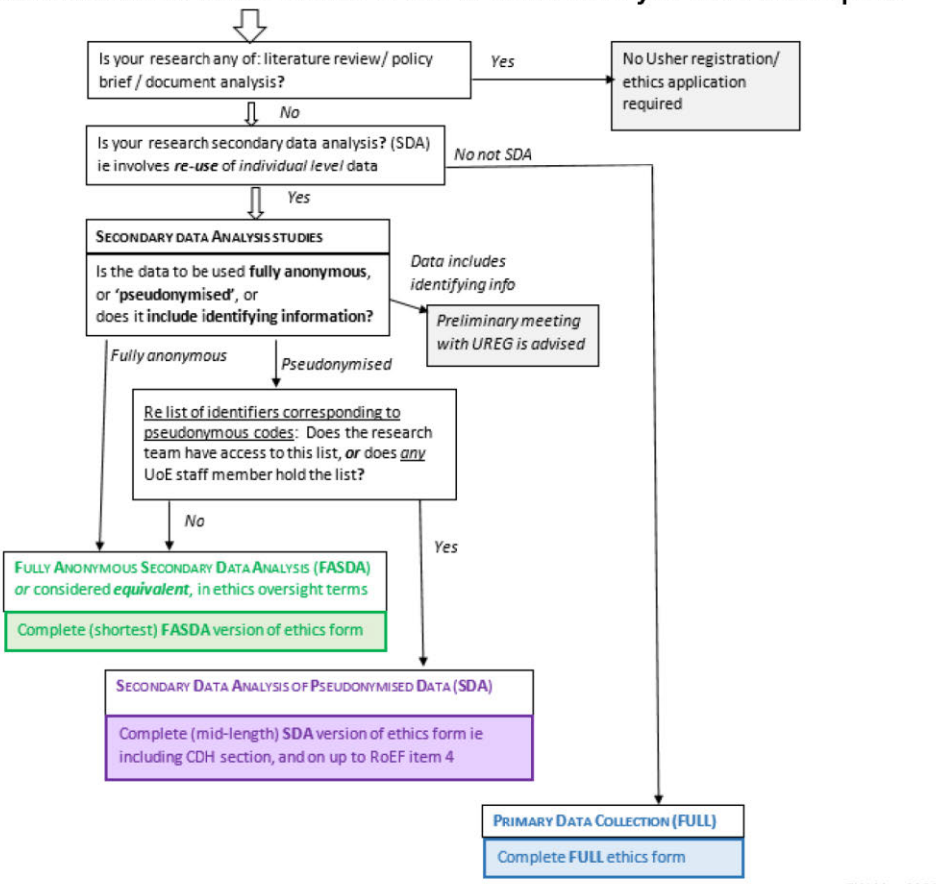
[1] *Path* U:\Datastore\CMVM\smgphs\shared\Usher\ResAdmin\ETHICSdocsForms

Appendix 7 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

UREG ETHICS APPLICATION FORM 'TRIAGE'

PROJECT REGISTRATION	*Please tick (✓) here to confirm that you are an **Usher Institute** member of staff/PG student →*	✓	
Name of Applicant:	Aryelly Rodriguez Carbonell		
Project Title (*this* sub-study/submission):	What is the risk of re-identification of individuals within publicly available clinical trial datasets?		
If applicant is a PG/UG student:			
Name(s) of Supervisor(s):	*Professor Steff Lewis, Professor Christopher Weir, Professor Sandra Eldridge and Dr Tracy Jackson*		
Please confirm supervisor concurs with triage outcome below	**Yes**		
Form indicated by use of 'triage' flow-chart below:	**FASDA**	**Date Triage undertaken:**	**03/DEC/2020**

Flow-chart to find which version of UREG ethics form you need to complete *



PW May 2020

* All 3 UREG ethics forms indicated above can be found at Usher Sharepoint or at path:
U:\Datastore\CMVM\smgphs\shared\Usher\ResAdmin\ETHICSdocsForms \CPHS Ethics Group docs\Forms to be completed

Appendix 7 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

THE UNIVERSITY of EDINBURGH | usher institute

USHER RESEARCH ETHICS GROUP (UREG)

GUIDANCE FOR ETHICS FORM: FULLY ANONYMOUS SECONDARY DATA ANALYSIS (FASDA)

This **FASDA** version of the UREG ethics form is to be completed by all staff and students conducting SDA of *fully-anonymised data*, as determined by use of the **'triage' flow chart.**

Suggested wording, for reporting completion of the **FASDA** version of the UREG ethics form, is:

"Prior to commencement, the research was subject to the Usher Institute ethics/data protection oversight process. The ethics/data protection triage and Overview self-audit of ethics/data protection issues, completed (by XXX°), confirmed that the proposed research (being YYY°) posed no reasonably foreseeable ethics/data protection risks. This indicated that there was no requirement for proceeding to full formal ethics/data protection review by the Usher Research Ethics Group."

Therefore completion of FASDA does *not* comprise 'ethics approval'. Nor even approval of data protection plans.

° *Where: XXX = 'the student and his/her academic supervisor', or 'the researcher/investigator'*
YYY = e.g. ' fully-anonymous secondary data analysis', or 'effectively… etc.

UoE Sponsorship

Your secondary data analysis research may require University of Edinburgh (ACCORD) **sponsorship** and/or **insurance**. New CMVM Accord policy is that *all* health or health-related 'research' studies using primary/ secondary data should in the first instance, *before application to UREG*, contact ACCORD re need or not for UoE sponsorship. *Please see doc* ***Research Studies within CMVM requiring UoE sponsorship****, at* [1] *and* Usher Sharepoint

Sponsorship is a *separate process* from ethics approval. Furthermore, after agreeing sponsorship, ACCORD might direct you to a *different* ethics committee ie *other than* UREG.

Note to PG/PGR student applicants and their supervisors:

The ethics form needs to be completed *under the oversight* of the PG/PGR student's UoE supervisor(s). For an online masters application there also needs to be oversight by the student's *'local'* supervisor. ***Responsibility for the validity and accuracy of the application lies with the academic supervisor(s).***

[1] *Path* U:\Datastore\CMVM\smgphs\shared\Usher\ResAdmin\ETHICSdocsForms\Other useful ethics docs\NHS R n D approval UoE Sponsorship

Appendix 7 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

UREG FORM FASDA (FULLY ANONYMOUS SECONDARY DATA ANALYSIS)

PROJECT REGISTRATION	*Please confirm (by ✓) that you are **Usher Institute** staff/PG student*	✓
Name of Applicant:		Aryelly Rodriguez Carbonell
Project Title (umbrella title for component(s) in *this* submission): If applicable: Titles of (sub) project(s) in *this* ethics application:		What is the risk of re-identification of individuals within publicly available clinical trial datasets?
Funding body *(if applicable):*		**Asthma UK Centre for Applied Research**
If applicant is a PG/UG student: **Degree for which registered:***		PhD in Medical Informatics
Deadline date for submission of work based on this research:		**31/JAN/2022**
If PGR degree, please also give (overall) topic of PhD research		**Are anonymised databases truly anonymous?**
Name(s) of Supervisor(s):		*Professor Steff Lewis, Professor Christopher Weir, Professor Sandra Eldridge and Dr Tracy Jackson*
Please confirm supervisor has thoroughly reviewed completion of form		**Yes**
Have you contacted ACCORD about UoE sponsorship of your research? *If you have not yet contacted ACCORD you **must** do so before submission to UREG*		**Yes**
If YES, *what was ACCORD's decision re UoE sponsorship?* *(delete one)*		UoE sponsorship: **Not needed**
If UoE sponsorship NEEDED		
(i) *please give date sponsorship was agreed (-in-principle)?*		**N/A**
(ii) *please confirm ACCORD letter submitted to UREG*		**N/A**
Date form submitted to UREG:		**03/12/2020**
Expected time-scale for the research (sub)study covered by *this* ethics application		**Data extract obtained: JAN 2021** **Date end this project: 31/JAN/2022**
List of (co-)investigators/ those involved in conducting the research, including names and positions *(e.g. PI, co-PI, res assoc, PhD student)*		Not applicable

* *eg campus MPH, online MPH, online MeH, UG Hons BMedSci with Usher supervisor, Usher-registered PGR [PhD, MRes, MPhil]*

Usher Ethics Form FASDA Aug 2020 v delta 1.5 page 2

Appendix 7 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

OVERVIEW (Self-audit)

The form should be completed by the principal investigator (PI) and if the PI is not a member of staff, it should be signed off by the PI's University supervisor.

Study type (please ✓ as appropriate *, and/or add a *new* type and ✓ that).

Tick one *only*	What type of research are you planning to do?
	Study utilising extract of routinely collected clinical data
	Meta-analysis utilising individual participant quantitative data collated from past studies
	Secondary data analysis within NHS or similar '**safe haven**' of linked pseudonymised data
✓	Secondary data analysis of data extract containing 'fully anonymous'[2] information
	Secondary data analysis of dataset containing ***not*** 'fully anonymous'[2] information
	Add in a new 'study type' descriptor if needed...

Please provide below a brief summary of your proposed study. *Not an entire protocol.*

Our interest here is ethical issues that might arise, so over and above research questions, study design and data protection, please focus on outlining/specifying:

- *Data owner (and institution), data sets to be linked if applicable and how, safe haven to be used if applicable, whether 'fully anonymous'[2] data or not, etc.*

Project Outline: *(word limit 400 words)*

There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. Some de-identified/anonymised dataset are now publicly available for secondary research. However, we do not know if they pose a privacy risk to the involved patients. We aim to collect a broad sample of these datasets in order to calculate their re-identification risks. Step 1: We will contact data holders and request access to their anonymised datasets following the data owners' local procedures. Step 2: Anonymised datasets will be explored at UoE and an overall re-identification risk will be calculated using two methodologies. Step 3: Reporting will be done on the obtained measures of risk without linking it to the obtained datasets. To the best of our knowledge, this will be the first study to calculate the risk of re-identification for these datasets using a unified strategy. For further details please refer to the study protocol (Version 1 dated 01DEC2020) which has been attached to this ethics form.

[2] For the purposes of classifying secondary data as 'fully anonymous', the following must both be true: (i) The data has had identifying features such as name, Chi number etc removed; (ii) If the data has been pseudonymised (with unique codes which are included in the data you receive for use), then you must ***not*** have for the data, **nor** have even *potential* access to, the identifying information corresponding to the codes, *and neither* must any other researcher in UoE even *hold* a list of identifiers for this secondary data.

Usher Ethics Form FASDA Aug 2020 v delta 1.5 page 3

Appendix 7 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

OVERVIEW *Please delete either **No** or **YES** answers as appropriate for your research project* ↘	**Overview audit responses**
1. Bringing the University into disrepute *(retain only answer that applies)* Is there any aspect of the proposed research that might bring the University into disrepute? For example, could any aspect of the research be considered controversial or prejudiced?	*No*
2. Given this is the 'FASDA' ethics form, your research must be secondary data analysis of 'fully anonymous' data from living individuals i.e. you do not have/ could not get access to, identifiable information for the data, *nor* does any other researcher anywhere in UoE have a list of identifiers for this secondary data. (If this is not the case please contact usher.ethics@ed.ac.uk)	
i. Are there any issues of DATA PROTECTION that are NOT adequately dealt with via established procedures? *(retain only answer that applies)* *A 'No' answer is justified (ie **YES** can be deleted) only if there is Caldicott Guardian, PBPP or other relevant/appropriate approval for your obtaining (access to) the secondary data and you intend to comply with all procedures specified, by the 'Data owner', for access to and use of that data.*	*No*
ii. Will you access UK Data Services, police, NHS or Social Care data on any living individual? ☐ tick ✓ if yes *If yes ✓ above, to 2 **ii**:* ~ Please confirm electronic attachment of Caldicott Guardian or similar approval for use of the data ☐ *tick if yes* ~ Did you complete a DPIA[3] in your application for access to that data? ☐ *tick if yes* If DPIA completed, is this completed DPIA attached herewith? ☐ *tick if yes*	*'Yes' ✓s within 2ii do not count towards Overview outcome*
3. Does your research fit into any of the following security-sensitive categories? If so, please indicate which by ticking all that apply. ☐ Commissioned by the military ☐ Commissioned under an EU security call ☐ Involve the acquisition of security clearances ☐ Concern groups which may be construed as terrorist or extremist **If you have *not* ticked any of these, please delete YES alongside**	*No*

Please list here ***external*** (non-UREG) approval/ DPIA documents being submitted, as indicated in **Overview** above:

OVERVIEW DOCUMENT SUBMISSION LIST Inventory of documentation promised in Overview section			
Document category	**Relating to item in Overview:**	**Tick (✓) if submitted**	**If submitted, name on electronic document**
Permission/approval for secondary data use	2 ii	NA	
'DPIA' for UKDS or NHS secondary data use	2 ii	NA	

[3] DPIA = DP Impact Assessment, but tick if DPIA *or equivalent* completed eg UKDS special Licence application

Usher Ethics Form FASDA Aug 2020 v delta 1.5 page 4

Appendix 7 for Evaluating Re-Identification Risks scores in Publicly Available Clinical Trial Datasets: Insights and Implications

TRIAGE AND OVERVIEW AUDIT OUTCOME:

Have you retained any YES answers *in last column*? *Please proceed as indicated below (➢):*

➢ *One or more* **Yes**es retained

Your responses on this completed self-audit indicate that **a degree of formal UREG ethics oversight is required.** You should NOT sign off below. Please contact UREG at usher.ethics@ed.ac.uk

➢ ***Not one*** **YES** retained

Your responses on the completed **Overview audit** confirm the ABSENCE OF REASONABLY FORESEEABLE ETHICS RISKS

Please sign off below - **ABSENCE OF REASONABLY FORESEEABLE ETHICAL or DATA PROTECTION RISK**

Usher Ethics Triage form, and FASDA Overview self-audit form, have been completed by the applicant for this research, and this has confirmed absence of reasonably foreseeable ethics or data protection risks. This indicates that formal ethics review by Usher Research Ethics Group is *not required*. Therefore, you will ***not*** receive an ethics 'approval' letter from UREG. However, if you require a UREG letter explaining this FASDA Overview self-audit within the Usher oversight process, you can request such in your covering email, or subsequently.

Applicant:	Aryelly Rodriguez Name	[redacted] Signature*	03DEC2020 Date
** **Supervisor** *(if student applicant)*	Professor Steff Lewis Name	[redacted] Signature	03Dec2020 Date

* Signatures or scanned in originals are acceptable.

** NOTE to Supervisor: The outcome of this **Overview** audit is based solely on the information/responses given in the 'triage' flowchart and in this form. In countersigning this Overview, as truly warranting no YES answers, you are taking responsibility, on behalf of the Usher Institute and the UoE, for the flowchart and ethics form completion).

Please submit this form in electronic format, together with Triage Flowchart form, and any additional docs listed as submitted in the box at the bottom of page 4, to usher.ethics@ed.ac.uk

For PG taught masters students, the submission should also be copied to pgadmin@ed.ac.uk

For PGR students, the submission should also be copied to s.georges@ed.ac.uk

Important: Please read these **Conditions for this Overview audit outcome**

If, subsequent to submission of this Form, there are *any* alterations to the outlined methods of the project, then the responses given in the Triage stage should be reviewed by the applicant and, if PG student, also by their supervisor. If the form identified has changed, then the new version should be completed for the revised study plan. If the triaged form is the same as before, then the responses on the form should be reviewed, to check that the Overview audit outcome is unchanged.

If a future change to data/methods results in a change to triage outcome, or to answers on the form, then **a resubmission of (an appropriate) form** is required, and it may then turn out that the Outcome then is different.

The principal investigator (and supervisor, if PI is a student) is/are responsible for ensuring compliance with any additional ethics requirements that might apply, and/or for ensuring compliance with any additional requirements for review by external bodies.

Usher Ethics Form FASDA Aug 2020 v delta 1.5 page 5

FINAL 1.0 23 July 2024 Page 7 of 7

File name: Appendix 7 ethical application 240723.docx

Appendix 4 - Chapter 4 - Datasets' metadata

<<<<<< REDACTED >>>>>>

Appendix 5 part 1 - Chapter 4 - Re-identification risk individual reports test

CONFIDENTIAL 1

Section 1. Unique Groups - 3 Indirect Identifiers

FaB 3 indirect identifiers		Sex	
		Male	Female
Centre	Age		
Edinburgh	<=70 years	12	86
	> 70 years	4	12
Sheffield	<=70 years	12	50
	> 70 years	1	8
North Bristol	<=70 years	5	43
	> 70 years	3	22
Newcastle	<=70 years		28
	> 70 years		2
Oxford	<=70 years	5	21
	> 70 years		10
Others	<=70 years	13	55
	> 70 years	3	18

FaB Risk Calculation Report - Draft 00 - tables run on: 11JUN2021 at 15:07
N = number of patients randomised, n = number of observations

By: Aryelly Rodriguez - ECTU Statistician

CONFIDENTIAL 2

Section 1. Unique Groups - 4 Indirect Identifiers

FaB 4 indirect identifiers			Smoking				
			>20 years and >15 cigarrets	>20 years	>15 cigarrets	Previously smoke	Never smoke
Centre	Age	Sex					
Edinburgh	<=70 years	Male		2		5	5
		Female	5	16		29	36
	> 70 years	Male				2	2
		Female				5	7
Sheffield	<=70 years	Male				6	6
		Female		4		19	27
	> 70 years	Male					1
		Female				3	5
North Bristol	<=70 years	Male	1	1		1	2
		Female	3	2		13	25
	> 70 years	Male	1			1	1
		Female				10	12
Newcastle	<=70 years	Female	1	1		5	21
	> 70 years	Female				1	1
Oxford	<=70 years	Male		1		2	2
		Female				12	9
	> 70 years	Female		1		2	7
Others	<=70 years	Male		2		7	4
		Female	2	1	1	19	32
	> 70 years	Male				1	2
		Female	1	2		6	9

FaB Risk Calculation Report - Draft 00 - tables run on: 11JUN2021 at 15:07
N = number of patients randomised, n = number of observations

By: Aryelly Rodriguez - ECTU Statistician

CONFIDENTIAL 3

Section 1. Unique Groups - 5 Indirect Identifiers

FaB 5 indirect identifiers			Smoking: >20 years and >15 cigarrets / Alcohol: <=14 units	Smoking: >20 years and >15 cigarrets / Alcohol: > 14 units	Smoking: >20 years / Alcohol: <=14 units	Smoking: >20 years / Alcohol: > 14 units	Smoking: >15 cigarrets / Alcohol: <=14 units	Smoking: Previously smoke / Alcohol: <=14 units	Smoking: Previously smoke / Alcohol: > 14 units	Smoking: Never smoke / Alcohol: <=14 units	Smoking: Never smoke / Alcohol: > 14 units
Centre	Age	Sex									
Edinburgh	<=70 years	Male				2		4	1	4	1
		Female	3	2	15	1		25	4	30	6
	> 70 years	Male						2		2	
		Female						4	1	7	
Sheffield	<=70 years	Male						3	3	2	4
		Female			3	1		18	1	26	1
	> 70 years	Male								1	
		Female						3		5	
North Bristol	<=70 years	Male	1		1			1		1	1
		Female	1	2	2			11	2	22	3
	> 70 years	Male	1						1	1	
		Female						9	1	12	
Newcastle	<=70 years	Female		1	1			5		17	4
	> 70 years	Female							1	1	
Oxford	<=70 years	Male			1			2		2	
		Female						9	3	7	2
	> 70 years	Female			1			2		7	
Others	<=70 years	Male			2			5	2	4	
		Female		2	1		1	16	3	31	1

(Continued)

FaB Risk Calculation Report - Draft 00 - tables run on: 11JUN2021 at 15:07 By: Aryelly Rodriguez - ECTU Statistician
N = number of patients randomised, n = number of observations

CONFIDENTIAL 4

Section 1. Unique Groups - 5 Indirect Identifiers

FaB 5 indirect identifiers			Smoking: >20 years and >15 cigarrets / Alcohol: <=14 units	Smoking: >20 years and >15 cigarrets / Alcohol: > 14 units	Smoking: >20 years / Alcohol: <=14 units	Smoking: >20 years / Alcohol: > 14 units	Smoking: >15 cigarrets / Alcohol: <=14 units	Smoking: Previously smoke / Alcohol: <=14 units	Smoking: Previously smoke / Alcohol: > 14 units	Smoking: Never smoke / Alcohol: <=14 units	Smoking: Never smoke / Alcohol: > 14 units
Centre	Age	Sex									
Others	> 70 years	Male						1		1	1
		Female	1		2			6		9	

FaB Risk Calculation Report - Draft 00 - tables run on: 11JUN2021 at 15:07 By: Aryelly Rodriguez - ECTU Statistician
N = number of patients randomised, n = number of observations

CONFIDENTIAL *5*

Section 2. Prosecutor Risks*

Number of Identifiers	*Parameter(s)*	*Tau level*	*Categories*	*Overall*
3	Entire dataset	.	All patients	421
		.	All unique levels	21
		.	Risk b (Rb)	1.000
		.	Risk c (Rc)	0.1727
	Over threshold tau 0.1	0.1	Levels above tau	8
		0.1	Patients on Levels above tau	31
		0.1	Risk a (Ra)	0.074
	Over threshold tau 0.2	0.2	Levels above tau	5
		0.2	Patients on Levels above tau	13
		0.2	Risk a (Ra)	0.031
	Over threshold tau 0.3	0.3	Levels above tau	4
		0.3	Patients on Levels above tau	9
		0.3	Risk a (Ra)	0.021
	Over threshold tau 0.4	0.4	Levels above tau	2
		0.4	Patients on Levels above tau	3
		0.4	Risk a (Ra)	0.007
4	Entire dataset	.	All patients	421
		.	All unique levels	60
		.	Risk b (Rb)	1.000
		.	Risk c (Rc)	0.4541
	Over threshold tau 0.1	0.1	Levels above tau	47
		0.1	Patients on Levels above tau	142
		0.1	Risk a (Ra)	0.337
	Over threshold tau 0.2	0.2	Levels above tau	33
		0.2	Patients on Levels above tau	55
		0.2	Risk a (Ra)	0.131
	Over threshold tau 0.3	0.3	Levels above tau	31
		0.3	Patients on Levels above tau	47
		0.3	Risk a (Ra)	0.112

FaB Risk Calculation Report - Draft 00 - tables run on: 11JUN2021 at 15:07 By: Aryelly Rodriguez - ECTU Statistician
*Using chapter 16 of El Emam, K. (2013). Guide to the De-Identification of Personal Health Information. New York: Auerbach Publications, https://doi.org/10.1201/b14764
Ra=The proportion of records that have a re-identification probability higher than a priori predetermined threshold tau
Rb=The maximum probability of re-identification in the dataset among all records
Rc=The proportion of records that can be correctly re-identified on average

CONFIDENTIAL *6*

Section 2. Prosecutor Risks*

Number of Identifiers	*Parameter(s)*	*Tau level*	*Categories*	*Overall*
	Over threshold tau 0.4	0.4	Levels above tau	29
		0.4	Patients on Levels above tau	41
		0.4	Risk a (Ra)	0.097
5	Entire dataset	.	All patients	421
		.	All unique levels	84
		.	Risk b (Rb)	1.000
		.	Risk c (Rc)	0.5443
	Over threshold tau 0.1	0.1	Levels above tau	73
		0.1	Patients on Levels above tau	190
		0.1	Risk a (Ra)	0.451
	Over threshold tau 0.2	0.2	Levels above tau	62
		0.2	Patients on Levels above tau	115
		0.2	Risk a (Ra)	0.273
	Over threshold tau 0.3	0.3	Levels above tau	55
		0.3	Patients on Levels above tau	87
		0.3	Risk a (Ra)	0.207
	Over threshold tau 0.4	0.4	Levels above tau	47
		0.4	Patients on Levels above tau	63
		0.4	Risk a (Ra)	0.150

FaB Risk Calculation Report - Draft 00 - tables run on: 11JUN2021 at 15:07 By: Aryelly Rodriguez - ECTU Statistician
*Using chapter 16 of El Emam, K. (2013). Guide to the De-Identification of Personal Health Information.
New York: Auerbach Publications, https://doi.org/10.1201/b14764
Ra=The proportion of records that have a re-identification probability higher than a priori predetermined threshold tau
Rb=The maximum probability of re-identification in the dataset among all records
Rc=The proportion of records that can be correctly re-identified on average

Appendix 5 part 2 - Chapter 4 - Re-identification risk individual reports for each dataset

<<<<<< REDACTED >>>>>>

Appendix 6 - Chapter 4 - Re-identification risk SAS code

Re-identification SAS core code

```
%MACRO RISK_CAL_PROSE (INDS, OUTDS, TPOPP, TAU_VAR);

PROC SORT DATA=&INDS; BY N; RUN;

DATA &INDS.1;
SET &INDS;
       IF _TYPE_='11111111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(8,6.)));
       IF _TYPE_='1111111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(7,6.)));
       IF _TYPE_='111111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(6,6.)));
       IF _TYPE_='11111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(5,6.)));
       IF _TYPE_='1111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(4,6.)));
       IF _TYPE_='111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(3,6.)));
       IF _TYPE_='11' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(2,6.)));
       TAU_THRESHOLD=&TAU_VAR;
        COUNT + 1;
        by N;
        IF FIRST.N THEN COUNT = 1;
       RISK_BY_CAT=1/N;
       IF RISK_BY_CAT > TAU_THRESHOLD THEN OVER_THRESHOLD=1;
       IF RISK_BY_CAT <= TAU_THRESHOLD THEN OVER_THRESHOLD=2;
       FORMAT OVER_THRESHOLD F01YN.;
RUN;

*---calculates summary stats;
%SUMSTAT(&INDS.1, &INDS.1_SUM, _TYPE_, RISK_BY_CAT, Entire dataset, 5, 1);

DATA &INDS.1_SUM;
       SET &INDS.1_SUM;
       N_IDENTIFY=&NUNNY;
       CVART=CVAR1*1;
RUN;

DATA &INDS.1_SUMRC;
       SET &INDS.1_SUM;
       IF STNAME="n";
       CVARTY=CVART/&TPOPP;
       SORT=3;
RUN;


DATA &INDS.2;
       SET &INDS.1;
       IF OVER_THRESHOLD=1;
RUN;

%let dsid = %sysfunc( open(&INDS.2) );
%let nobs = %sysfunc( attrn(&dsid,nobs) );
%let rc = %sysfunc( close(&dsid) );

%IF &nobs=0 %THEN %DO; *put dummy template to indicate dataset does not have observations;

       PROC IMPORT
           DATAFILE= "&ROOT\NO_TAUP.xlsx"
           OUT= WORK.&INDS.2_SUM REPLACE
              DBMS=xlsx ;
           sheet="sheet1";
       RUN;

       DATA &INDS.2_SUM;
       SET &INDS.2_SUM;
       N_IDENTIFY=&NUNNY;
       TAU_THRESHOLD=&TAU_VAR;
       VARLABEL="Over threshold tau &TAU_VAR";
       CVART=0;
       CVARTY=0;
       RUN;

%END;

%IF &nobs>0 %THEN %DO;
```

SAS code – re-identification Page 1 of 5

Re-identification SAS core code

```
*---SUMSTAT calculates basic summary stats;
%SUMSTAT(&INDS.2, &INDS.2_SUM, _TYPE_, N, Over threshold tau &TAU_VAR, 5, 2);

DATA &INDS.2_SUM;
       SET &INDS.2_SUM;
       N_IDENTIFY=&NUNNY;
       TAU_THRESHOLD=&TAU_VAR;
       CVART=CVAR1*1;
       CVARTY=CVART/&TPOPP;
RUN;

%END;

DATA INTERIM ;
       SET &INDS.1_SUM &INDS.1_SUMRC &INDS.2_SUM;
       IF STATID=8 AND SORT=3 THEN RISKID=1; *Rc;
       IF STATID=5 AND SORT=1 THEN RISKID=2; *Rb;
       IF STATID=10 AND SORT=2 THEN RISKID=3; *Ra;

       IF STATID=8 AND SORT=1 THEN RISKID=81; *All levels;
       IF STATID=8 AND SORT=2 THEN RISKID=91; *Levels where threshold applied;

       IF STATID=9 AND SORT=1 THEN RISKID=82; *All patients;

       IF RISKID NE . ;

       IF STATID=10 AND SORT=2 THEN CVAR1=PUT(CVARTY,9.3); *;

       IF STATID=9 AND SORT=1 THEN DO; CVAR1=PUT(&TPOPP,9.0); CVART=&TPOPP; END; *All patients;

RUN;

DATA &INDS.2_SUM2;
       SET &INDS.2_SUM;
       IF STATID=10 AND SORT=2; *All patients where threshold applies;
       RISKID=92;
RUN;

DATA &OUTDS (DROP=STNAME CVART CVARTY STLABEL);
       SET INTERIM &INDS.2_SUM2;
       %ALIGN(CVAR1);
       IF RISKID IN (1) THEN CVAR1=PUT(CVARTY,12.5);
       IF RISKID IN (2) THEN CVAR1=PUT(CVART,12.5);
       IF RISKID IN (3) THEN CVAR1=PUT(CVARTY,12.5);
       IF RISKID NOT IN (1, 2, 3) THEN CVAR1=PUT(CVART,12.0);
       FORMAT RISKID N01RISK.;
RUN;

PROC SORT DATA=&OUTDS; BY RISKID; RUN;

*---clean environment;
PROC DATASETS LIBRARY=WORK NOLIST;
      DELETE INTERIM &INDS.1 &INDS.2 &INDS.1_SUM &INDS.2_SUM &INDS.2_SUM2 ;
QUIT;

%MEND RISK_CAL_PROSE;


%MACRO RISK_CAL_JOU (INDS, OUTDS, TPOPP, TAU_VAR);

%SUMSTAT(RISKRAW_T, &INDS.5_SUM, _TYPE_, N_ORI, Entire dataset, 5, 5);

PROC SORT DATA=&INDS; BY N; RUN;

%SUMSTAT(&INDS, &INDS._SUM, _TYPE_, N, Entire dataset, 5,  0);

DATA &INDS.1;
       SET &INDS;
       IF _TYPE_='11111111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(8,6.)));
       IF _TYPE_='1111111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(7,6.)));
       IF _TYPE_='111111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(6,6.)));
```

SAS code – re-identification Page 2 of 5

Re-identification SAS core code

```
      IF _TYPE_='11111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(5,6.)));
      IF _TYPE_='1111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(4,6.)));
      IF _TYPE_='111' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(3,6.)));
      IF _TYPE_='11' THEN CALL SYMPUT("NUNNY",COMPRESS(PUT(2,6.)));
      TAU_THRESHOLD=&TAU_VAR;
      INDI_RISK=N_ORI/N; *---fj/Fj;
      RISK_BY_CAT=1/N; *---1/Fj=1/N;
      IF INDI_RISK > 1 THEN INDI_RISK=1; *---correction due to imprecision of syn data;
      IF INDI_RISK > TAU_THRESHOLD THEN OVER_THRESHOLD=1;
      IF INDI_RISK <= TAU_THRESHOLD THEN OVER_THRESHOLD=2;
      IF N_ORI NE .;
      FORMAT OVER_THRESHOLD F01YN.;
      LABEL INDI_RISK="fj/Fj individual class risk"  RISK_BY_CAT="1/Fj";
RUN;

%SUMSTAT(&INDS.1, &INDS.3_SUM, _TYPE_, RISK_BY_CAT, Entire dataset, 5, 3); * max of 1/Fj;

%SUMSTAT(&INDS.1, &INDS.10_SUM, _TYPE_, INDI_RISK, Entire dataset, 5, 10); * max fj/Fj, it needs
divided by n in original dataset;

%SUMSTAT(&INDS.1, &INDS.11_SUM, _TYPE_, N, Entire dataset, 0, 11);

DATA &INDS.2;
SET &INDS.1;
IF OVER_THRESHOLD=1;
RUN;

%let dsid = %sysfunc( open(&INDS.2) );
%let nobs = %sysfunc( attrn(&dsid,nobs) );
%let rc = %sysfunc( close(&dsid) );

%IF &nobs=0 %THEN %DO; *put dummy template to indicate dataset does not have observations;
PROC IMPORT
    DATAFILE= "&ROOT\NO_TAUJ.xlsx"
    OUT= WORK.&INDS.2_SUM REPLACE
       DBMS=xlsx ;
    sheet="sheet1";
RUN;

DATA &INDS.2_SUM;
      SET &INDS.2_SUM;
      TAU_THRESHOLD=&TAU_VAR;
RUN;

%END;


%IF &nobs > 0 %THEN %DO;
      %SUMSTAT(&INDS.2, &INDS.2_SUM, _TYPE_, N_ORI, Over threshold tau &TAU_VAR, 3, 2);
      *---8,2 Y 10,2 ;
      DATA &INDS.2_SUM;
      SET &INDS.2_SUM;
      TAU_THRESHOLD=&TAU_VAR;
      RUN;
%END;

DATA NBOOBS (KEEP=DUMMY NNN_SYN);
      SET &INDS.11_SUM;
      IF STATID=10 AND SORT=11; *N observations used from syn dataset;
      DUMMY=1;
      NNN_SYN=CVAR1*1;
RUN;

DATA Nn (KEEP=DUMMY NNN_ORI);
      SET &INDS.5_SUM;
      IF STATID=10 AND SORT=5; *n All patients in original dataset;
      DUMMY=1;
      NNN_ORI=CVAR1*1;
RUN;

DATA INTERIM1 ;
```

Re-identification SAS core code

```
      SET &INDS.5_SUM &INDS._SUM &INDS.3_SUM &INDS.10_SUM &INDS.11_SUM &INDS.2_SUM ;
      N_IDENTIFY=&NUNNY;
      CVART=CVAR1*1;

      IF STATID=10 AND SORT=5 THEN RISKID=83; *n All patients in original dataset;
      IF STATID=8 AND SORT=5 THEN RISKID=84;  *All unique levels in original dataset;

      IF STATID=10 AND SORT=0 THEN RISKID=81;  *All observations in syn dataset;
      IF STATID=8 AND SORT=0 THEN RISKID=82;  *All unique levels syn dataset;

      IF STATID=10 AND SORT=11 THEN RISKID=85; *N observations used from syn dataset;

      IF STATID=5 AND SORT=3 THEN RISKID=2;  *Risk b (Rb);

      IF STATID=8 AND SORT=2 THEN RISKID=86; *Levels above tau;

      IF STATID=10 AND SORT=2 THEN RISKID=87;  *Patients on levels above tau;

      IF RISKID NE . ;
RUN;

DATA INTER ;
      SET &INDS.5_SUM &INDS.10_SUM  &INDS.2_SUM ;
      N_IDENTIFY=&NUNNY;
      CVART=CVAR1*1;
      IF STATID=8 AND SORT=5 THEN DO; RISKID=1.2; CVART=(CVAR1*1); END; *Risk c2 (Rc) it needs
      divided by N in syn dataset of selected;
      IF STATID=10 AND SORT=10 THEN DO; RISKID=1.1; CVART=(CVAR1*1); END; *Risk c1 (Rc) it needs
      divided by n in original dataset;
      IF STATID=10 AND SORT=2 THEN DO; RISKID=3; CVART=(CVAR1*1); END; *Patients on levels above
      tau, dividing this by n gives Ra;
      IF RISKID NE . ;
      DUMMY=1;
RUN;

%KEEP_1ST(INTER,NBOOBS,DUMMY,AAA);

%KEEP_1ST(AAA,Nn,DUMMY,INTERIM2);

DATA INTERIM;
      SET INTERIM1 INTERIM2;
      IF RISKID=1.2 THEN CVART=CVART/NNN_SYN;
      IF RISKID=1.1 THEN CVART=CVART/NNN_ORI;
      IF RISKID=3  THEN CVART=CVART/NNN_ORI;

      IF CVART=. AND RISKID=3 THEN DO; CVART=0; VARLABEL="Over threshold tau &TAU_VAR"; END;
      IF CVART=. AND RISKID IN (86, 87) THEN DO; CVART=0; VARLABEL="Over threshold tau &TAU_VAR";
      END;
RUN;

PROC SORT DATA=INTERIM; BY RISKID; RUN;

DATA INTERIM;
      SET INTERIM;
      IF RISKID IN (1.1, 1.2, 2, 3) AND CVART>=1 THEN CVART=1;
RUN;

DATA INTERIM;
      SET INTERIM;
      %ALIGN(CVAR1);
RUN;

DATA &OUTDS (DROP=STNAME CVART STLABEL);
      SET INTERIM ;
      IF RISKID IN (1.1, 1.2, 2, 3) THEN CVAR1=PUT(CVART,12.5);
      IF RISKID NOT IN (1.1, 1.2, 2, 3) THEN CVAR1=PUT(CVART,12.0);
      %ALIGN(CVAR1);
      FORMAT RISKID N02RISK.;
RUN;

PROC SORT DATA=&OUTDS; BY RISKID; RUN;
```

Re-identification SAS core code

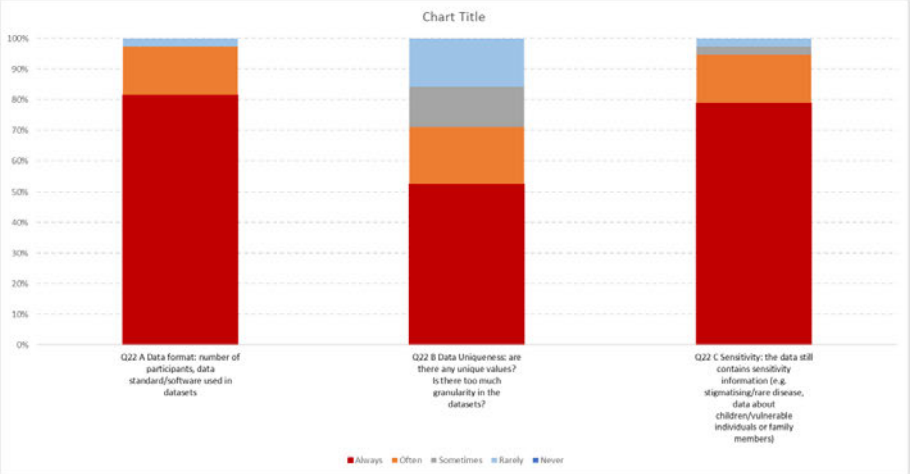
```
*---clean environment;
PROC DATASETS LIBRARY=WORK NOLIST;
      DELETE  &INDS.1 &INDS.3_SUM &INDS.5_SUM &INDS.10_SUM &INDS.11_SUM &INDS.2 &INDS.2_SUM
&INDS._SUM  INTERIM INTERIM1 INTERIM2 AAA NBOOBS Nn INTER  ;
QUIT;

%MEND RISK_CAL_JOU;

*---Prosecutor scenario;

*---upload dataset;
ods select none; *---option to turn off in case it is disclosive;
*---Create strata;
PROC TABULATE DATA=DATASET FORMAT=10. OUT=DATASET_RISKRAW ;
       CLASS AGE SEX EDUCATION OCCUPATION;
       TABLE AGE='Age'*SEX='Sex'*n=' ',EDUCATION='Education'*OCCUPATION='Occupation'
       /BOX="&STUDY_UNIID"
       ROW=FLOAT
       MISSTEXT=' ';
RUN;
ods select all;

%RISK_CAL_PROSE (DATASET_RISKRAW, DATASET_RISK_001,&TPOP, 0.01);
%RISK_CAL_PROSE (DATASET_RISKRAW, DATASET_RISK_005,&TPOP, 0.05);
%RISK_CAL_PROSE (DATASET_RISKRAW, DATASET_RISK_01, &TPOP, 0.1);
%RISK_CAL_PROSE (DATASET_RISKRAW, DATASET_RISK_02, &TPOP, 0.2);
%RISK_CAL_PROSE (DATASET_RISKRAW, DATASET_RISK_03, &TPOP, 0.3);
%RISK_CAL_PROSE (DATASET_RISKRAW, DATASET_RISK_04, &TPOP, 0.4);
%RISK_CAL_PROSE (DATASET_RISKRAW, DATASET_RISK_05, &TPOP, 0.5);


*---Journalist scenario;

*---upload original dataset;
ods select none; *---option to turn off in case it is disclosive;
*---Create strata from original dataset;
PROC TABULATE DATA=DATASET FORMAT=10. OUT=DATASET_RISKRAW;
       CLASS AGE SEX EDUCATION OCCUPATION;
       TABLE AGE='Age'*SEX='Sex'*n=' ',EDUCATION='Education'*OCCUPATION='Occupation'
       /BOX="&STUDY_UNIID"
       ROW=FLOAT
       MISSTEXT=' ';
RUN;
ods select all;

*---upload synthetic dataset;
ods select none;
*---Create strata from synthetic dataset;
PROC TABULATE DATA=Syn_DATASET FORMAT=10. OUT= RISKRAW_SYN ;
       CLASS AGE SEX EDUCATION OCCUPATION;
       TABLE AGE='Age'*SEX='Sex'*n=' ',EDUCATION='Education'*OCCUPATION='Occupation'
       /BOX="&STUDY_UNIID"
       ROW=FLOAT
       MISSTEXT=' ';
RUN;
ods select all;

*---Match strata from original dataset with synthetic dataset;
%KEEP_1ST (RISKRAW_SYN, RISKRAW, IDMATCH, DATASET_JRISK);

%RISK_CAL_JOU (DATASET_JRISK, DATASET_JRISK_001,&TPOP, 0.01);
%RISK_CAL_JOU (DATASET_JRISK, DATASET_JRISK_005,&TPOP, 0.05);
%RISK_CAL_JOU (DATASET_JRISK, DATASET_JRISK_01, &TPOP, 0.1);
%RISK_CAL_JOU (DATASET_JRISK, DATASET_JRISK_02, &TPOP, 0.2);
%RISK_CAL_JOU (DATASET_JRISK, DATASET_JRISK_03, &TPOP, 0.3);
%RISK_CAL_JOU (DATASET_JRISK, DATASET_JRISK_04, &TPOP, 0.4);
%RISK_CAL_JOU (DATASET_JRISK, DATASET_JRISK_05, &TPOP, 0.5);
```

SAS code – re-identification Page 5 of 5

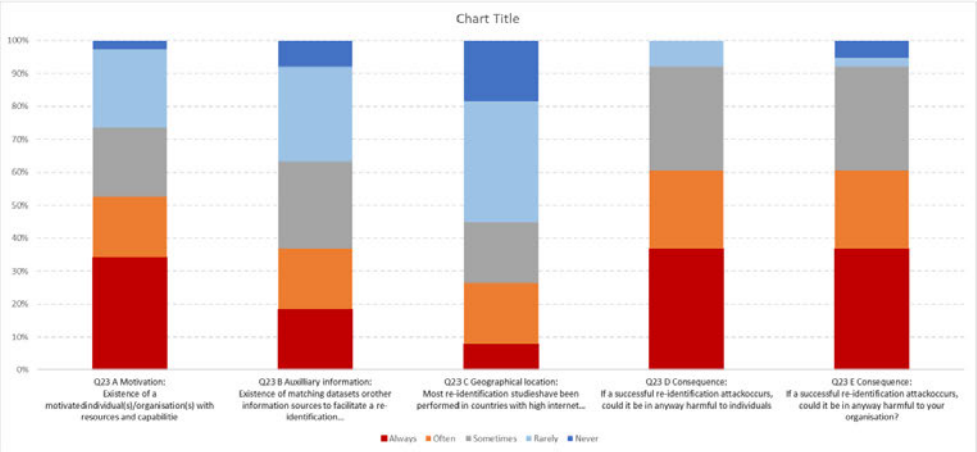
Appendix 7 - Chapter 5 - Published supplementary materials.

Supplementary Figures

Supplementary Figure 1 Awareness of risk parameters associated with the dataset



Supplementary Figure 2 Awareness of risk parameters associated with the environment



Additional file 1 - Study protocol

What are the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets? A study protocol

Rodriguez A[1], Lewis SC[1], Eldridge S[2], Jackson T[3], Weir CJ[1]

[1]Edinburgh Clinical Trials Unit, Usher Institute, the University of Edinburgh

[2]Pragmatic Clinical Trials Unit, Blizard Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London

[3]Asthma UK Centre for Applied Research, Usher Institute, the University of Edinburgh

Correspondence:

Ms Aryelly Rodriguez

Edinburgh Clinical Trials Unit, the University of Edinburgh

Level 2, Nine Edinburgh BioQuarter,

9 Little France Road, Edinburgh, EH16 4UX

Emails:

What are the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets? A study protocol

Abstract

There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. However there is no a single standardised set of recommendations on how to anonymise and prepare clinical trial datasets for sharing and an ever increasing number of anonymised clinical trials datasets are becoming available for secondary research. Therefore, this study aims to explore the current views and experiences of researchers in the UK about de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets.

Key Words: Clinical Trials | Data Anonymisation | Re-identification | De-identification | Datasets | Re-identification risk

What are the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets? A study protocol

29 **Definitions**

Anonymisation	A data set would be considered anonymised if it has been de-identified and then subsequent data manipulation/steps have been taken to further protect the dataset, for example, if a privacy model has been applied (e.g. k-anonymity) or the link with the original non anonymised dataset has been destroyed and this action cannot be reversed.
De-identification	Removal of all personal health information and all other indirect identifiers which could lead to the identification of an individual. The most common de-identification methods are: 1. HIPPA (US Health Insurance Portability and Accountability Act of 1996) Safe harbour, in which 18 identifiers are removed from the datasets [1] [2] 2. Hrynaszkiewicz et al. [3] proposed an enhanced removal of potential identifiers which are commonly present in clinical trials datasets.
Controlled Access	Datasets that can only be accessed if permission is granted by the data holders via their internal procedures.
Open Access:	Datasets that can be accessed without any or minimal restrictions imposed by the data holders.
Publicly available datasets	Data sets that are discoverable and available for sharing via open or controlled access, this data can be located on central repositories or with individual institutions/researchers
Re-identification risk score	Estimated probability of any given individual being re-identified from an anonymised/de-identified dataset. The re-identification risk score depends on the variables available in the dataset, the number of observations in the dataset and on the strategy used to attack the dataset (prosecutor or journalist scenario).
Prosecutor scenario	If the adversary knows that a target individual (for whom identifiers are known) is in the publicly available dataset (released anonymised and/or de-identified) we are under prosecutor re-identification risk scores. This scenario seeks to identify uniqueness in the records of the publicly available dataset.
Journalist scenario	If the adversary sets out to identify any individual from the publicly available dataset just to prove that it can be done by using another dataset for "matching" with the publicly available dataset, then we are under journalist re-identification risk scores.
Secondary Research	Consist of using already existing data for addressing questions out of scope for the original research which collected the data (primary research).

30

31

What are the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets? A study protocol

Background

There is now a strong drive, particularly from publishers and funders, to encourage the release of relevant anonymised trial data sets [4]. Therefore, data-sharing has become an essential activity to disseminate current research, to enable new investigations and to maximise the scientific endeavour [5] [6] . Currently there are a number of such anonymised datasets made publicly available for secondary research via open or controlled access [7] [8]. Anonymisation of data is complex, and its implementation often means that the detail necessary to fully analyse the data is lost. There is therefore a balance between wanting to de-risk a dataset prior to sharing, against wanting it to be sufficiently detailed to answer valid research questions. So, we propose to explore the United Kingdom (UK) researchers' views regarding their experiences with the creation and release of de-identified/anonymised clinical trial datasets, the generation and use of re-identification risk scores, and their views about wider aspects of re-identification risks.

Why it is important to do this study?

We are currently investigating the re-identification risk scores across a range of clinical trials datasets [9]. Therefore, we want to better understand the views of UK researchers regarding their experiences with the creation and release of de-identified/anonymised clinical trial datasets, the generation and use of re-identification risk scores, and their views about wider aspects of re-identification risks. Knowing how other researchers are using such scores and in which context will help us identify how they could be useful in the future.

What are the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets? A study protocol

Objective

To explore clinical trials researchers' views on their experiences with the creation and release of de-identified/anonymised clinical trial datasets, the generation and use of re-identification risk scores, and the wider aspects of re-identification risks.

Methods

Survey Design

The "checklist of questions for designing a survey study plan" by Creswell et at [10] was followed for the development of this protocol (see Appendix 1).

This study will use an online exploratory cross-sectional descriptive survey [10] [11] that consists of both open-ended and close-ended questions for data collection. This will allow us to gather information to better describe actual experiences regarding the investigated topic. The open-ended questions are especially important because of the lack of previous reporting on researchers' views and experiences.

The survey will be in English. Most of the close-ended questions will have mutually exclusive choices, with a smaller number allowing for multiple answers [12] [13]. Where applicable, close-ended questions, will have an "other" (free text) option added to allow participants to provide an answer that is not available for selection [12]. Five-point response scales will be used for questions assessing frequency (always, often, sometimes, rarely, never).

The survey will be structured in five parts:

1. Consent and eligibility check.
2. Researchers' work background details (current position, years of experience in current position and general place of work)

3 Researchers' experiences with the creation and release of de-identified/ anonymised clinical trial datasets

4. Researchers' awareness, knowledge and use regarding the generation of re-identification risk scores

5. Researchers' views about wider aspects of re-identification risks

The first draft of the survey is presented in Appendix 2 of this protocol.

The survey is designed to follow the layout presented in Figure 1.

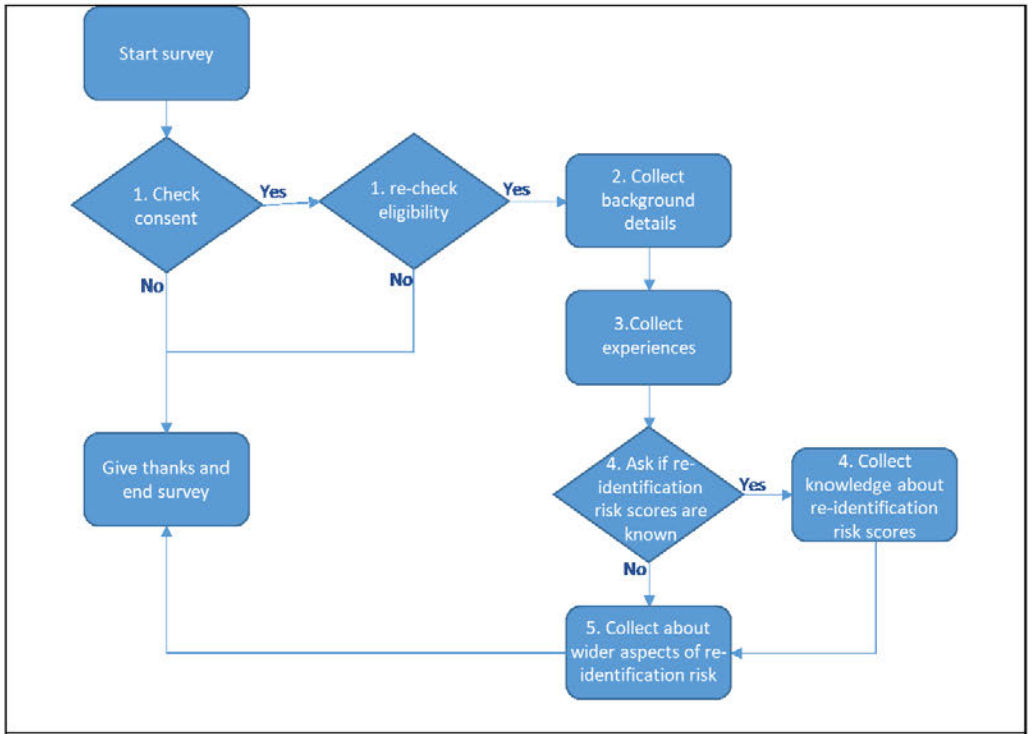


Figure 1. Single survey data collection process.

Therefore, a single participant (after the eligibility criteria has been met) will answer between 17 and 22 questions out of the proposed 24 questions, as some answers will determine the relevance of the next question.

The survey will be piloted with a selection of The University of Edinburgh personnel with experience in the processes of de-identification /anonymisation, release/maintenance and/or re-identification risk assessment of clinical trial datasets in order to prepare them for secondary research, before it is finalised and sent to the intended participants.

Study Population

Inclusion/Exclusion criteria: Clinical trial researchers based in the UK with experience in executing/overseeing any of the processes of de-identification /anonymisation, release/maintenance and/or re-identification risk assessment of clinical trial datasets in order to prepare them for secondary research.

Sampling and Recruitment

There will not be a formal sample size or stratification of the surveyed researchers as this is an exploratory study. Therefore, we will use convenience non-probability sampling[11] [14] [15] by providing a MS Forms[16] link or QR code with an introduction (email or print-out) (see Appendix 3) to:

- All 52 Clinical Trial Units (CTUs) registered in the UKCRC network (https://ukcrc-ctu.org.uk/registered-ctus/). We will email all UK fully/provisionally registered CTUs. We expect to obtain at least one survey per CTU. (population size n=52)
- The data transparency group at PHUSE (https://phuse.global/) (Contact via email, population size unknown).
- Allstat@JISCMAIL.AC.UK, an email discussion list for the UK Education and Research communities. (https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=allstat) (Contact via email, population size unknown).
- Participants at the 6th International Clinical Trials Methodology Conference (ICTMC) (3-6 October 2022) (Special event) (Contact via leaflet, Population size unknown).

The aim is to obtain as many responses as possible while the survey is active (around 5 weeks) to maximise the range of experiences. We estimate the population to be heterogeneous so a minimum of between 12-30 surveys is required[14] to reflect a wide range of views.

Data collection and extraction

This survey will not collect any personal data from the clinical trial researchers, and after extraction, all open questions will be carefully checked to make sure their coding do not contain any identifiable information. Only AR will be able to access all the data. We will use MS Forms as it provides the integrated web interface and data collection for the survey. The data by MS Forms "is encrypted both at rest and in transit" and it is stored in a European Server, all compliant with the General Data Protection Regulation (GPDR), for more detail please see:

https://support.microsoft.com/en-us/office/security-and-privacy-in-microsoft-forms-7e57f9ba-4aeb-4b1b-9e21-b75318532cd9

https://support.microsoft.com/en-us/office/data-storage-for-microsoft-forms-97a34e2e-98e1-4dc2-b6b4-7a8444cb1dc3

Analysis.

When the active period for the survey ends the response summary information and the individual responses of the complete surveys will be exported from MS Forms directly to AR's secured and password protected area at UoE or AR's DataStore allocation as per University of Edinburgh data handling policies. [17-19].

Individual responses will be kept until December 2023, then destroyed in accordance with University of Edinburgh policy for destroying archived research data (see https://www.ed.ac.uk/sites/default/files/atoms/files/dataprotectionhandbookv9.pdf and https://www.ed.ac.uk/data-protection/data-protection-policy).

Close-ended questions will be analysed using descriptive statistics (counts and percentages) on SAS 9.4 (or a more recent version). All this data will be analysed by AR and sense checked by another investigator (SCW, CJW, TJ).

Thematic analysis [20] [21] will be used to generate themes from the open-ended questions responded using NVivo® 12 (or a more recent version). All the data will be coded by AR and themes will be reviewed, defined and finalised in discussion with the multi-disciplinary research team to ensure valuable perspectives were included and help reduce subjectivity of findings (SCW, CJW, TJ).

This survey cannot investigate the response rate or the response bias to the survey. The results of this study will help to understand the views of UK researchers regarding their experiences with the creation and release of de-identified/anonymised clinical trial datasets, the generation and use of re-identification risk scores, and their views about wider aspects of re-identification risks.

153 **Timetable**

154 This is the proposed timetable for this study.

	Months					
	May2022	Jun2022	Sep2022	Oct2022	Nov2022	Dec2022
Ethics & Governance						
Protocol						
Pilot survey						
Main survey						
Analysis						
Report						

155

156 **Potential sources of bias and limitations**

157 This study is covering an emerging part of clinical trials research, for which even consensus
158 about the definition of anonymisation does not exist [22]. This will create a variety in the views
159 and experiences on de-identification / anonymisation of clinical trials datasets.
160 We acknowledge that this survey will potentially be filled out by highly motivated individuals with
161 positives experiences, and we might not be fully engaging with researchers who have done de-
162 identification / anonymisation but have had adverse experiences or difficulties in this area.
163 As this topic is very dense and our resources are limited, we are going to base the survey in the
164 UK. Therefore it will not be possible to explore what is happening in other countries.

165

166 **Ethics and dissemination**

167 This project will not collect identifiable or personal participant data or personal sensitive
168 information; therefore, this is a low risk project and we will follow the ethical review
169 processes coordinated by the Edinburgh Medical School Research Ethics Committee
170 (EMREC). Findings from this research will be presented at scientific conferences and

published in a peer-reviewed journal. No publication or presentation originating from this work will reveal any data that could lead to re-identification of individuals from the data collected.

Conflicts of Interests

The authors declare no competing interests.

Funding

AR has a scholarship from the University of Edinburgh to undertake a PhD with support from the Asthma UK Centre for Applied Research (AUKCAR). Neither funder (University of Edinburgh) nor sponsor (AUKCAR) contributed to protocol development.

CJW is supported in this work by NHS Lothian via the Edinburgh Clinical Trials Unit.

SCL is supported in this work by her employment at the Edinburgh Clinical Trials Unit.

TJ is supported by Asthma UK as part of the Asthma UK Centre for Applied Research (grant nos. AUK-AC-2012-01 and AUK-AC-2018-01),

SE is supported in this work by her employment at the Pragmatic Clinical Trials Unit.

Author contributions

AR, SCL and CJW conceived the idea for this work supported by SE. AR wrote the first draft, and all authors contributed to this article.

192 ***Appendix 1 A checklist of Questions for Designing a Survey Study Plan***

193 ***(extracted from Chapter 8 in Research Design by John Creswell and J. David***

194 ***Creswell, 5th edition.) [10]***

195

Item ID	Item description	Protocol compliance
1	Is the purpose of the survey stated?	Yes
2	Are the reasons for choosing the design mentioned?	Yes
3	Is the nature of the survey (cross-sectional vs. longitudinal) identified	Yes
4	Is the population and its size mentioned?	Yes
5	Will the population be stratified? If so how?	Yes
6	How many people will be in the sample? On what basis was this size chosen?	Yes
7	What will be the procedure for sampling these individuals (e.g. random, non-random)?	Yes
8	What instrument will be used in the survey? Who developed the instrument?	Not applicable
9	What are the content areas addressed in the survey? The Scales	Yes
10	What procedure will be used to pilot or field test the survey?	Yes
11	What is the timeline for administering the survey?	Yes
12	What are the variables in the survey?	Yes
13	How do these variables cross-reference with the research questions and items on the survey?	Yes
14	What specific steps will be taken in data analysis to do the following	
14a	Analyse returns?	Yes
14b	Check for response bias?	Yes
14c	Conduct a descriptive analysis?	Yes
14d	Collapse items into scales?	Yes
14e	Check for reliability of scales?	Not applicable
14f	Run inferential statistics to answer research questions or assess practical implications of the results?	Not applicable
15	How will the results be interpreted?	Yes

196

197

198

199 ***Appendix 2 Survey (currently a separate file)***

200 ***Appendix 3 Introduction email (currently a separate file)***

201

Reference List

1. U.S. Government, *Health Insurance Portability and Accountability Act of 1996*, in *Public law*. 1996. p. 191.
2. U.S. Department of Health & Human Services (HHS), *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. 2012.
3. Hrynaszkiewicz, I., et al., *Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers*. Trials, 2010. **11**(340).
4. Dal-Ré, R., *Access to Anonymized Individual Participant Clinical Trials Data: A Radical Change of Mind by the Most Prestigious Medical Journals*. Archivos de Bronconeumologia, 2018. **54**(2): p. 65-67.
5. Pisani, E., et al., *Beyond open data: realising the health benefits of sharing data*. BMJ, 2016. **355**: p. i5295.
6. Bertagnolli, M., et al., *Advantages of a truly open-access data-sharing model*. N Engl J Med, 2017. **12**(376): p. 1178-1181.
7. Clinical Study Data Request (CSDR). *Clinical Study Data Request*. Available from: https://clinicalstudydatarequest.com/.
8. The Yale University. *Yale University Open Data Access (YODA) Project*. [cited 2020 26 Oct 2020]; Available from: http://yoda.yale.edu/.
9. Rodriguez, A., et al., *What are the re-identification risk scores of publicly available anonymised clinical trial datasets? A study protocol* 2020, The University of Edinburgh p. 19.
10. Creswell, J.W. and J.D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. 5th ed. 2018: Sage publications.
11. Fink, A., *The survey handbook*. 2003: sage.
12. Boynton, P.M. and T. Greenhalgh, *Selecting, designing, and developing your questionnaire*. Bmj, 2004. **328**(7451): p. 1312-1315.
13. Stehr-Green, P.A., et al., *Developing a questionnaire*. FOCUS Field Epidemiol, 2003. **2**(2): p. 1-6.
14. Dudovskiy, J., *The ultimate guide to writing a dissertation in business studies: A step-by-step assistance*. Pittsburgh, USA, 2016: p. 51.
15. Lavrakas, P.J., *Encyclopedia of survey research methods*. 2008, Thousand Oaks, California: Sage publications.
16. Microsoft, *MS Forms*. 2016. p. Part of Office 365.
17. The University of Edinburgh. *Data - Data Services*. 2020 [cited 2020 30 Oct 2020].
18. The University of Edinburgh. *Use University services*. 2020 [cited 2020 30 Oct 2020]; Available from: https://www.ed.ac.uk/infosec/information-protection-policies/procedures-guidance/use-university-services.
19. The University of Edinburgh. *Working with sensitive data* 2020 [cited 2020 30 Oct 2020]; Available from: https://www.ed.ac.uk/information-services/research-support/research-data-service/during/sensitive-data.
20. Braun, V. and V. Clarke, *Using thematic analysis in psychology*. Qualitative research in psychology, 2006. **3**(2): p. 77-101.
21. Gibbs, G.R., *Thematic coding and categorizing*. Analyzing qualitative data, 2007. **703**: p. 38-56.
22. Rodriguez, A., et al., *Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review*. 2022, The Univeristy of Edinburgh.

Additional file 2 - Survey questions

An online survey of UK researchers' views on sharing anonymised clinical trials datasets.

Thank you for responding to this academic survey as part of a PhD at the Usher Institute at the University of Edinburgh. We would be grateful for your participation and value your views and experiences about de-identification, anonymisation, data release and re-identification risk estimation processes related to clinical trials datasets. The answers obtained will enhance the available evidence on this topic. Answering the survey questions should take no more than 15 minutes.

* Required

Participant Information about the survey

Introduction:
There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. However there is no a single standardised set of recommendations on how to anonymise and prepare clinical trial datasets for sharing. Therefore, this survey aims to explore the current views and experiences of researchers in the UK about de-identification/anonymisation, release methods and re-identification risk estimation processes for clinical trials datasets.

Study description:
This short questionnaire is arranged in four sections. The first section seeks information about your role. The second section collects your experience with the creation and release of de-identified/anonymised clinical trial datasets. The third section is about your views regarding the generation and use of re-identification risk scores. Finally, the fourth section is about your views about wider aspects of re-identification risks. Text boxes in key questions allow you to add your comments, please do not leave them blank, if they are not applicable, just write N.A. The questions have been carefully selected and piloted to capture relevant views about the subject. The responses to most of the questions have been streamlined to require only a few clicks to select the answer. Answering the survey will take about 15 minutes to complete.

Risks and benefits:
Participation in this study is voluntary. You can withdraw from the study at any point before you have submitted your responses without giving a reason. The Edinburgh Medical School Research Ethics Committee (EMREC at The University of Edinburgh) has approved the survey (Approval Reference: XXXXX).
The responses to the survey will be anonymised at source and stored on a secured server provided by the University of Edinburgh. There are no benefits associated with your participation in this survey, however, this academic research could help to improve the knowledge about the existing practices on de-identification/anonymisation, data release and re-identification risk estimation.
Should the survey results be published, the anonymity of the responses will be not compromised. Further information about the Usher Institute's Privacy policy is available at http://www.ed.ac.uk/files/atoms/files/usher_institute_-_privacy_notice_v2.0_0.pdf (http://www.ed.ac.uk/files/atoms/files/usher_institute_-_privacy_notice_v2.0_0.pdf)

Eligibility and invitation for participation:
We would appreciate your participation in this survey. Your opinion is very important to us. To proceed to the survey, please read the participant consent form and if satisfied, give your consent below.

1. Participant Consent Form

I agree to take part in this study and confirm the following:

- I have read the information provided above about this survey.
- I have had the opportunity to consider the information, which has been sufficient for giving my consent and I understand I will not receive any payment for taking part.
- I understand that my participation is voluntary and that I am free to withdraw at any time by closing the browser window without giving any reason and without my legal rights being affected. However, once I have pressed the submit button, I cannot withdraw my data as it is anonymous.
- I understand that the data collected during the study will be analysed by individuals at the Usher Institute at the University of Edinburgh and regulatory authorities for audit purposes.
- I understand that taking part involves providing anonymous survey data and the responses may be shared with other researchers and used in future research projects, published in journals, and be used for teaching or academic material.
- By checking the box below, I confirm that I have understood and agree with the above statements, and I consent to taking part in this study.

*

◯ I confirm I have read the above and give my consent.

Section 1. Background

For this survey:
De-identification refers to the removal of all personal health information and all other indirect identifiers which could lead to the identification of an individual. The most common de-identification methods are HIPPA (US Health Insurance Portability and Accountability Act of 1996) Safe harbour, in which 18 identifiers are removed from the datasets (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html {https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html}) and Hrynaszkiewicz et al. (2010) (https://www.bmj.com/content/340/bmj.c181 {https://www.bmj.com/content/340/bmj.c181}) with an enhanced removal of potential identifiers which are commonly present in clinical trials datasets.
Anonymisation is when a dataset has been de-identified and then subsequent data manipulation/steps have been taken to further protect the dataset, for example, if a privacy model has been applied (e.g. k-anonymity) or the link with the original non anonymised dataset has been destroyed and this action cannot be reversed.
Secondary analysis: the use of existing de-identified/anonymised datasets, which were collected for a prior study, to investigate research questions outside the scope of the original study.

2. Eligibility. Are you involved or have you ever been involved with the de-identification/anonymisation of clinical trials dataset and/or with the processes for releasing de-identified/anonymised clinical trials dataset for secondary analysis? *

- ◯ Yes
- ◯ No
- ◯ Maybe

3. Are you based in the United Kingdom (UK)? *

- ◯ Yes
- ◯ No

4. Place of work. Are you currently affiliated to, or working at? *

- ◯ an UKCRC (Fully or provisionally) Registered Unit
- ◯ [ ] Other

5. What is your current job title or role? (Please choose the closest) *

- ◯ Director / Senior Manager
- ◯ Principal investigator
- ◯ Researcher (Research fellow/associate)
- ◯ Senior Researcher
- ◯ Statistician
- ◯ Senior Statistician
- ◯ Trial manager / coordinator
- ◯ Senior Trial manager / coordinator
- ◯ [ ] Other

6. Years of experience in that role * 3 / 11

- ◯ 0 to 2
- ◯ 3 to 5
- ◯ 6 or 10
- ◯ More than 10

Section 2. Your experiences on creation/release of de-identified/anonymised datasets

For this survey:
De-identification refers to the removal of all personal health information and all other indirect identifiers which could lead to the identification of an individual. The most common de-identification methods are HIPPA (US Health Insurance Portability and Accountability Act of 1996) Safe harbour, in which 18 identifiers are removed from the datasets (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html)) and Hrynaszkiewicz et al. (2010) (https://www.bmj.com/content/340/bmj.c181 (https://www.bmj.com/content/340/bmj.c181)) with an enhanced removal of potential identifiers which are commonly present in clinical trials datasets.
Anonymisation is when a dataset has been de-identified and then subsequent data manipulation/steps have been taken to further protect the dataset, for example, if a privacy model has been applied (e.g. k-anonymity) or the link with the original non anonymised dataset has been destroyed and this action cannot be reversed.
Data release under controlled access: Datasets that can only be accessed if permission is granted by the data holders via their internal procedures.
Data release under open access: Datasets that can be accessed without any or minimal restrictions imposed by the data holders.

7. How are/were you involved with the de-identification/anonymisation of clinical trials dataset and/or with the processes for releasing de-identified/anonymised clinical trials dataset? (please tick all that apply) *

- [ ] Creation/Generation of de-identified/anonymised dataset
- [ ] Evaluation/Assessment/Peer review of de-identified/anonymised dataset
- [ ] Approval of the release of de-identified/anonymised dataset
- [ ] Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk
- [ ] Uploading/Maintenance/Distribution of de-identified/anonymised dataset
- [ ] Other

8. Years of experience in dealing with de-identification/anonymisation and/or release of clinical trial datasets *

- ◯ 0 to 2
- ◯ 3 to 5
- ◯ 6 or 10
- ◯ More than 10

9. Which kind of documents do you currently use to assist you with the process to prepare and release de-identified/anonymised clinical trial datasets? *

- ◯ Only internally developed documents/guidance (e.g. standardised operating procedures (SOPs), work instructions)
- ◯ Only externally sourced documents/guidance (e.g. MRC guidance https://www.methodologyhubs.mrc.ac.uk/files/7114/3682/3831/Datasharingguidance2015.pdf (https://www.methodologyhubs.mrc.ac.uk/files/7114/3682/3831/Datasharingguidance2015.pdf), ICO anonymisation guidelines https://ico.org.uk/media/1061/anonymisation-code.pdf (https://ico.org.uk/media/1061/anonymisation-code.pdf))
- ◯ Both internal and external documents/guidance
- ◯ Other

10. The internally developed documents/guidance (policies, SOPs or working instructions) at your organisation help to guide you 5 / 11 with:

	Yes (documents/guidance implemented)	Yes (but document/guidance under construction)	No (this process is not covered)
The process of how to de-identify/anonymise clinical trials datasets	○	○	○
The process for releasing de-identified/ anonymised clinical trials datasets	○	○	○
The assessment of the re-identification risk of the de-identified/ anonymised clinical trials datasets.	○	○	○

11. Which process do you mostly use for releasing de-identified/anonymised clinical trial data *

- ○ Only de-identification, under controlled access
- ○ De-identification followed by anonymisation, under controlled access
- ○ Only de-identification, under open access
- ○ De-identification followed by anonymisation, under open access
- ○ Other

12. In your opinion, how was your experience in the process of **de-identifying/anonymising** clinical trials datasets? (e.g. what have worked? what have not worked? any concerns? any assurances?) *

13. In your opinion, how was your experience in the process of **releasing** de-identified/anonymised clinical trials datasets? (e.g. what have worked? what have not worked? any concerns? any assurances?) *

Version 01 - 28 April 2022 RodriguezA_risk_researchers_opinion_survey_final_01_220428.pdf

4/28/2022

14. In your opinion, how was your experience in the process of **maintaining** released de-identified/anonymised clinical trials datasets? (e.g. what have worked? what have not worked? any concerns? any assurances?) * 6 / 11

Section 3. Your views about re-identification risks scores

For this survey:
Re-identification risk scores are defined as the estimated probabilities of any given individual being re-identified from an anonymised/de-identified dataset. The re-identification risk score depends on the variables available in the dataset, the number of observations in the dataset and on the strategy used to attack the dataset (prosecutor or journalist scenario).
Prosecutor scenario is when the adversary knows that a target individual (for whom identifiers are known) is in the publicly available dataset (released anonymised and/or de-identified).
Journalist scenario is when the adversary sets out to identify any individual from the publicly available dataset just to prove that it can be done by using another dataset for "matching" with the publicly available dataset.
(for more information view chapter 13 of El Emam, K. (2013). *Guide to the de-identification of personal health information*. CRC Press)

15. Are you aware of or ever come across re-identification risk scores for assisting in the release of de-identified/anonymised clinical trials datasets? *

- ◯ I have never heard of them
- ◯ I have heard about them, but I am not so sure what they are
- ◯ I have a general understanding, but do not use them
- ◯ I have a good understanding, and use them sometimes
- ◯ I have a strong understanding, and use them frequently

16. Which scenario(s) do you consider for calculating the re-identification risk score? *

- ◯ Prosecutor scenario (the adversary knows that a target individual is in the publicly available dataset)
- ◯ Journalist scenario (the adversary sets out to identify any individual from the publicly available dataset)
- ◯ Both
- ◯ Other: _____

17. Please, describe how the calculated re-identification risk scores inform the process for data release. *

_____

18. Which program do you use for calculating the re-identification risk scores? *

- ◯ SAS
- ◯ CRAN - Package sdcMicro
- ◯ Other: _____

19. In your opinion, what are the advantages/disadvantages of using re-identification risk scores? * 8 / 11

20. How likely are you going to continue to use/generate the re-identification risk score? *

0	1	2	3	4	5	6	7	8	9	10

Not at all likely Extremely likely

Section 3. Your views about re-identification risks score (Cont.)

For this survey:
Re-identification risk scores are the estimated probability of any given individual being re-identified from an anonymised/de-identified dataset. The re-identification risk score depends on the variables available in the dataset, the number of observations in the dataset and on the strategy used to attack the dataset (prosecutor or journalist scenario).

21. What are the barrier for not using the re-identification risk scores (please tick all that apply) *

- [ ] Lack of funding
- [ ] Lack of relevant training
- [ ] Lack of time
- [ ] Other

Section 4. Your views about wider aspects of re-identification risk

22. Each anonymised clinical trial dataset is unique, but in general terms, before any release of an anonymised dataset, do you think about: *

	Always	Often	Sometimes	Rarely	Never
Data format: number of participants, data standard/software used in datasets	○	○	○	○	○
Data Uniqueness: are there any unique values? Is there too much granularity in the datasets?	○	○	○	○	○
Sensitivity: the data still contains sensitivity information (e.g. stigmatising/rare disease, data about children/vulnerable individuals or family members)	○	○	○	○	○

23. Each anonymised clinical trial dataset is surrounded by a unique environment, but in general terms before any release of an anonymised dataset, do you think about: *

	Always	Often	Sometimes	Rarely	Never
Motivation: Existence of a motivated individual(s)/organisation(s) with resources and capabilities for an re-identification attack	○	○	○	○	○
Auxilliary information: Existence of matching datasets or other information sources to facilitate a re-identification attack	○	○	○	○	○
Geographical location: Most re-identification studies have been performed in countries with high internet penetration, arguably because there are more data available about individuals to make such attempts more likely to succeed.	○	○	○	○	○
Consequence: If a successful re-identification attack occurs, could it be in anyway harmful to individuals?	○	○	○	○	○
Consequence: If a successful re-identification attack occurs, could it be in anyway harmful to your organisation?	○	○	○	○	○

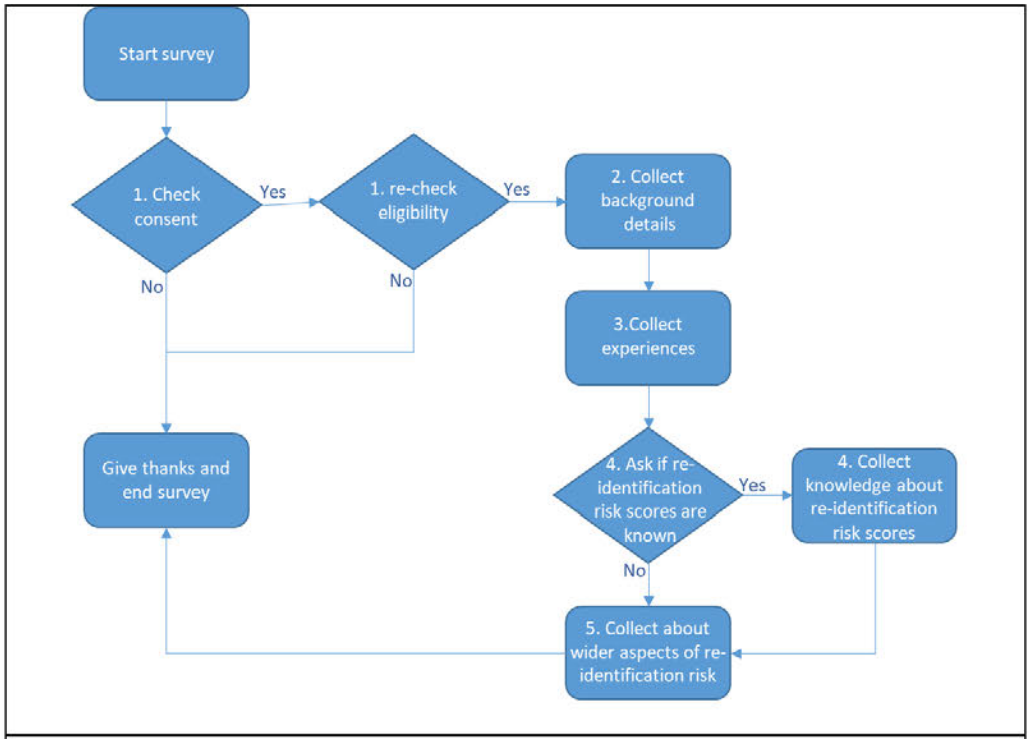
24. Please write below any other aspects you consider before you release de-identified /anonymised clinical trials data. * 11 / 11

Additional file 3 A checklist of Questions for Designing a Survey Study Plan

(Extracted from Chapter 8 in Research Design by John Creswell and J. David Creswell, 5th edition.)

Item ID	Item description	Protocol compliance
1	Is the purpose of the survey stated?	Yes
2	Are the reasons for choosing the design mentioned?	Yes
3	Is the nature of the survey (cross-sectional vs. longitudinal) identified	Yes
4	Is the population and its size mentioned?	Yes
5	Will the population be stratified? If so how?	Yes
6	How many people will be in the sample? On what basis was this size chosen?	Yes
7	What will be the procedure for sampling these individuals (e.g. random, non-random)?	Yes
8	What instrument will be used in the survey? Who developed the instrument?	Not applicable
9	What are the content areas addressed in the survey? The Scales	Yes
10	What procedure will be used to pilot or field test the survey?	Yes
11	What is the timeline for administering the survey?	Yes
12	What are the variables in the survey?	Yes
13	How do these variables cross-reference with the research questions and items on the survey?	Yes
14	What specific steps will be taken in data analysis to do the following	
14a	Analyse returns?	Yes
14b	Check for response bias?	Yes
14c	Conduct a descriptive analysis?	Yes
14d	Collapse items into scales?	Yes
14e	Check for reliability of scales?	Not applicable
14f	Run inferential statistics to answer research questions or assess practical implications of the results?	Not applicable
15	How will the results be interpreted?	Yes

Additional file 4 Survey data collection process.



Survey data collection process.

Additional file 5 Invitation (email or print-out) for survey

An online survey of UK researchers' views on sharing anonymised clinical trials datasets. - Survey closing on the 19th October 2022

There are increasing pressures for anonymised datasets from clinical trials to be shared across the scientific community. However there is no a single standardised set of recommendations on how to anonymise and prepare clinical trial datasets for sharing and an ever increasing number of anonymised clinical trials datasets are becoming available for secondary research.

As part of my PhD research, I'm working with my supervisors the University of Edinburgh to examine this topic and would like to hear the views and experiences of researchers in the UK about de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets.

It would be very useful to hear your views. The survey is available until the 19th October 2022. Please use the link or the QR code for more details about this survey and thank you!

https://forms.office.com/r/0aTLCVqmE3

The survey will take about 15 minutes to complete and all responses will be anonymised and kept confidential.

Should you have any further questions, please contact me at:

Aryelly Rodriguez

Statistician/PhD candidate, Edinburgh Clinical Trials Unit, The University of Edinburgh, UK.

Additional file 6 - List of registered CTUs (13NOV2022)

igniting our potential

UKCRC Registered Clinical Trials Units

2022/23 UKCRC Registration ID Numbers

ID	ORGANISATION	STATUS
32	Barts and the London Pragmatic CTU	FULL
4	Barts Clinical Trials Unit	FULL
1	Birmingham Clinical Trials Unit	FULL
70	Bristol Trials Centre*	FULL
3	CaCTUS (Cancer Clinical Trials Unit Scotland)	FULL
55	Cambridge Clinical Trials Unit (CCTU)	FULL
64	Cambridge Epidemiology & Trials Unit	FULL
6	Cancer Research UK Clinical Trials Unit (CRCTU)	FULL
7	Centre for Healthcare Randomised Trials (CHaRT)	FULL
63	Centre for Trials Research	FULL
56	Comprehensive CTU @ UCL	FULL
5	CR UK & UCL Cancer Trials Centre	FULL
14	Diabetes Trials Unit (Churchill Hospital, Oxford)	FULL
67	Derby Clinical Trials Support Unit (DCTSU)	FULL
15	Edinburgh Clinical Trials Unit, Edinburgh	FULL
65	Exeter Clinical Trials Unit	FULL
16	Glasgow Clinical Trials Unit	FULL
18	Imperial Clinical Trials Unit	FULL
42	Intensive Care National Audit & Research Centre (ICNARC) CTU	FULL
36	Keele Clinical Trials Unit	FULL
53	King's Clinical Trials Unit at King's Health Partners	FULL
41	Leeds Clinical Trials Research Unit	FULL
43	Leicester Clinical Trials Unit	FULL
12	Liverpool Trials Collaborative	FULL
44	London School of Hygiene & Tropical Medicine	FULL
9	Manchester Clinical Trials Unit	FULL
19	Medical Research Council Clinical Trials Unit at UCL	FULL
22	Newcastle Clinical Trials Unit (NCTU)	FULL
57	NHS Blood and Transplant Clinical Trials Unit	FULL
23	North Wales Organisation for Randomised Trials in Health (NWORTH)	FULL
25	Northern Ireland Clinical Trials Unit	FULL
51	Norwich Clinical Trials Unit	FULL
26	Nottingham Clinical Trials Unit	FULL
21	NPEU Clinical Trials Unit	FULL
46	Oxford Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU)	FULL
27	Oxford Clinical Trials Research Unit (OCTRU)	FULL
52	Oxford Primary Care and Vaccines Collaborative Clinical Trials Unit	FULL
60	Papworth Trials Unit Collaboration	FULL
31	Peninsula Clinical Trials Unit	FULL
20	PRIMENT Clinical Trials Unit at UCL	FULL
62	Royal Marsden Clinical Trials Unit (RM-CTU)	FULL
34	Sheffield Clinical Trials Research Unit	FULL
37	Southampton Clinical Trials Unit	FULL
61	Surrey Clinical Trials Unit	FULL
58	Swansea Trials Unit	FULL
49	Tayside Clinical Trials Unit	FULL
17	The Institute of Cancer Research Clinical Trials & Statistics Unit (ICR- CTSU)	FULL

2022/23 UKCRC Registration ID Numbers

ID	ORGANISATION	STATUS
39	Warwick Clinical Trials Unit	FULL
40	York Trials Unit	FULL
68	Hull Health Trials Unit	PROV
69	Lancashire Clinical Trials Unit	PROV
66	Brighton and Sussex Clinical Trials Unit	PROV – UNDER REVIEW

*** Bristol Trials Centre is a merger of the Bristol Clinical Trials & Evaluation Unit (Reg ID 11) and the Bristol Randomised Trials Collaboration (Reg ID 2) both of which were fully registered units.**

Additional file 7 - Favourable opinion on ethics

THE UNIVERSITY of EDINBURGH

Edinburgh Medical School
Research Ethics Committee (EMREC)

emrec@ed.ac.uk

Aryelly Rodriguez-Carbonell
Clinical Trials Statistician
Edinburgh Clinical Trials Unit

01 June 2022

Dear Aryelly

Study Title: What are the UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trials datasets

REC Reference: 22-EMREC-027

The Research Ethics Committee has now reviewed the above application.

Ethical opinion

The Committee can give a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, with no conditions.

In your EMREC form, the answer to E1 was missing; we have made the relevant amendment and saved with file name "EMREC Ethics Form v05 AIR_v 1.1 0106".

Amendments and Reporting Requirements

Now that you have a favourable ethical opinion from EMREC you are bound to the protocol, informed consent and data collection materials reviewed by us. Small changes like updating contact details, fixing typos, or adding a partner's logo do not require an amendment. However, you must re-contact us if you wish to make substantive changes that affect the protocol or answers to any of the questions on the EMREC form.

You should also contact EMREC to notify us about:

- Serious breaches of the protocol
- Safety reports and any adverse events

Favourable opinion from EMREC is not the only requirement to ensure integrity in research conduct, so in parallel with this stage, you will also want to satisfy yourself that you have considered other governance issues.

Documents reviewed

THE UNIVERSITY of EDINBURGH

The final list of documents reviewed and approved by the Committee is as follows:

22-EMREC-027 Please quote this number on all correspondence		
Document	*Version*	*Date*
EMREC Ethics Form	1.1	01/06/2022
Cover letter	1.0	04/05/2022
Protocol	1.0	28/04/2022
Survey (including information and consent form)	1.0	28/04/2022
Survey email	1.0	04/05/2022
ACCORD email (sponsorship not required)	1.0	29/04/2022
DPIA	1.0	03/05/2022
HRA tool results	1.0	28/04/2022
Data protection certificate	1.0	28/03/2020
DP for Research certificate	1.0	28/04/2022

With the Committee's best wishes for the success of this project.

Yours sincerely,

Sue Fletcher-Watson
Co-Chair, EMREC

Christine Campbell
Co-Chair, EMREC

Additional file 8 - Survey Closed Questions dataset

ID	Duration in survey	Email	Q1 Participant Consent Form	Q2 Eligibility	Q3 Are you based in the United Kingdom (UK)?	Q4 Place of work. Are you currently affiliated to, or working at	Q5 What is your current job title or role? (Please choose the closest)	Q6 Years of experience in that role	Q7 How are, were you involved
1	00:04:54	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	More than 10	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;
2	00:27:24	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Director / Senior Manager	More than 10	Approval of the release of de-identified/anonymised dataset;Creation/Generation of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;
3	00:08:34	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Professor / Line manager	6 or 10	Creation/Generation of de-identified/anonymised dataset;
4	00:15:59	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Statistician	More than 10	Creation/Generation of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
5	00:00:36	anonymous	I confirm I have read the above and give my consent.	No					
6	00:00:33	anonymous	I confirm I have read the above and give my consent.	No					
7	00:00:17	anonymous	I confirm I have read the above and give my consent.	No					
8	00:06:41	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	Other	Principal Investigator	More than 10	Approval of the release of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;
9	00:15:44	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	Other	Retired	More than 10	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;
10	00:13:45	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	Other	Professor and Chair of Statistics	6 or 10	Evaluation/Assessment/Peer review of de-identified/anonymised dataset;
11	00:01:03	anonymous	I confirm I have read the above and give my consent.	No					
12	00:10:14	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Researcher (Research fellow/associate)	6 or 10	Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
13	00:08:27	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	6 or 10	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
14	00:00:45	anonymous	I confirm I have read the above and give my consent.	No					
15	00:07:13	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	3 to 5	Creation/Generation of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
16	00:17:17	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Governance facilitator	3 to 5	Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;Creation/Generation of de-identified/anonymised dataset;
17	00:11:08	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Director / Senior Manager	More than 10	Approval of the release of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;
18	00:01:38	anonymous	I confirm I have read the above and give my consent.	No					
19	00:00:16	anonymous	I confirm I have read the above and give my consent.	No					
20	00:42:53	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Director / Senior Manager	3 to 5	Approval of the release of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;
21	00:04:01	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Principal Investigator	3 to 5	Creation/Generation of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;
22	00:49:25	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Researcher (Research fellow/associate)	3 to 5	Creation/Generation of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;Creation of SOPs and guidelines for de-identifying and releasing datasets;
23	00:04:36	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Trial manager / coordinator	0 to 2	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
24	02:26:20	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Head of IT	3 to 5	Creation/Generation of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
25	01:06:09	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	Other	Senior Researcher	More than 10	Not directly involved;
26	00:01:22	anonymous	I confirm I have read the above and give my consent.	No					
27	00:08:25	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	More than 10	Creation/Generation of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
28	00:00:55	anonymous	I confirm I have read the above and give my consent.	No					
29	00:07:50	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Director / Senior Manager	6 or 10	Creation/Generation of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;
30	00:10:23	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	More than 10	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
31	00:06:45	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Principal Investigator	More than 10	Creation/Generation of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;
32	00:02:12	anonymous	I confirm I have read the above and give my consent.	No					
33	00:21:56	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Statistician	6 or 10	Creation/Generation of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
34	00:01:29	anonymous	I confirm I have read the above and give my consent.	No					
35	00:37:53	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Statistician	0 to 2	Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Creation/Generation of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;
36	00:00:09	anonymous	I confirm I have read the above and give my consent.	No					
37	00:24:09	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	3 to 5	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
38	02:07:17	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	More than 10	Approval of the release of de-identified/anonymised dataset;
39	00:16:56	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Data Manager/IS	More than 10	Creation/Generation of de-identified/anonymised dataset;
40	00:20:23	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Trial manager / coordinator	0 to 2	Evaluation/Assessment/Peer review of de-identified/anonymised dataset;
41	00:05:25	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	Other	Principal Investigator	More than 10	Approval of the release of de-identified/anonymised dataset;
42	00:08:25	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Statistician	6 or 10	Creation/Generation of de-identified/anonymised dataset;
43	00:05:02	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	More than 10	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;
44	00:05:44	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Director / Senior Manager	More than 10	Approval of the release of de-identified/anonymised dataset;
45	00:05:18	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	Other	Senior Statistician	6 or 10	Uploading/Maintenance/Distribution of de-identified/anonymised dataset;Creation/Generation of de-identified/anonymised dataset;
46	00:03:46	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Director / Senior Manager	3 to 5	Creation/Generation of de-identified/anonymised dataset;Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Generation/Evaluation/Assessment of de-identified/anonymised dataset re-identification risk;
47	00:01:02	anonymous	I confirm I have read the above and give my consent.	No					
48	00:14:04	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Statistician	More than 10	Creation/Generation of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;
49	00:00:47	anonymous	I confirm I have read the above and give my consent.	No					
50	00:12:45	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Statistician	6 or 10	Creation/Generation of de-identified/anonymised dataset;
51	01:53:55	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Senior Statistician	More than 10	Evaluation/Assessment/Peer review of de-identified/anonymised dataset;Approval of the release of de-identified/anonymised dataset;
52	00:10:08	anonymous	I confirm I have read the above and give my consent.	Yes	Yes	an UKCRC (fully or provisionally) Registered Unit	Statistician	More than 10	Creation/Generation of de-identified/anonymised dataset;Uploading/Maintenance/Distribution of de-identified/anonymised dataset;

ID	Q8 Years of experience in dealing with	Q9 Which kind of documents do you currently use to assist you	Q10A The process of how to de-identify_anonymise	Q10B The process for releasing	Q10C The assessment of the re-identification ris	Q11 Which process do you mostly use for releasing
1	3 to 5	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)		De-identification followed by anonymisation, under controlled access
2	More than 10	Both internal and external documents/guidance	No (this process is not covered)	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	Only de-identification, under controlled access
3	6 or 10	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
4	3 to 5	Both internal and external documents/guidance	No (this process is not covered)	No (this process is not covered)	No (this process is not covered)	Only de-identification, under controlled access
5						
6						
7						
8	0 to 2	Only externally sourced documents/guidance				Only de-identification, under controlled access
9	6 or 10	Only internally developed documents/guidance	No (this process is not covered)	No (this process is not covered)	No (this process is not covered)	Only de-identification, under open access
10	More than 10	Only internally developed documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
11						
12	3 to 5	Only internally developed documents/guidance				Not applicable
13	6 or 10	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
14						
15	0 to 2	Both internal and external documents/guidance	Yes (but document/guidance under construction)	Yes (but document/guidance under construction)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
16	0 to 2	Both internal and external documents/guidance	Yes (but document/guidance under construction)	Yes (but document/guidance under construction)	No (this process is not covered)	Only de-identification, under controlled access
17	0 to 2	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (but document/guidance under construction)	Yes (documents/guidance implemented)	Only de-identification, under controlled access
18						
19						
20	0 to 2	Only internally developed documents/guidance	Yes (but document/guidance under construction)	No (this process is not covered)	Yes (but document/guidance under construction)	De-identification followed by anonymisation, under controlled access
21	3 to 5	None				Only de-identification, under controlled access
22	3 to 5	Both internal and external documents/guidance	Yes (but document/guidance under construction)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by some aspects of anonymisation,***redacted***
23	3 to 5	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (but document/guidance under construction)	Yes (but document/guidance under construction)	De-identification followed by anonymisation, under controlled access
24	More than 10	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	Yes (but document/guidance under construction)	De-identification followed by anonymisation, under controlled access
25	0 to 2	Not applicable				Not applicable
26						
27	More than 10	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	Yes (but document/guidance under construction)	Only de-identification, under controlled access
28						
29	3 to 5	Both internal and external documents/guidance	Yes (but document/guidance under construction)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
30	0 to 2	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	De-identification followed by anonymisation, under controlled access
31	6 or 10	Other				Only de-identification, under controlled access
32						
33	0 to 2	Only externally sourced documents/guidance				De-identification followed by anonymisation, under controlled access
34						
35	0 to 2	Only internally developed documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	Only de-identification, under controlled access
36						
37	0 to 2	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	No (this process is not covered)	Only de-identification, under controlled access
38	3 to 5	Only internally developed documents/guidance	Yes (documents/guidance implemented)			De-identification followed by anonymisation, under controlled access
39	More than 10	Both internal and external documents/guidance	Yes (but document/guidance under construction)	Yes (documents/guidance implemented)	Yes (but document/guidance under construction)	Only de-identification, under controlled access
40	0 to 2	Other				***redacted***
41	6 or 10	Only externally sourced documents/guidance				De-identification followed by anonymisation, under controlled access
42	3 to 5	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	No (this process is not covered)	Only de-identification, under controlled access
43	3 to 5	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
44	6 or 10	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	Yes (but document/guidance under construction)	***redacted***
45	0 to 2	Only internally developed documents/guidance	Yes (but document/guidance under construction)	Yes (but document/guidance under construction)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
46	3 to 5	Only internally developed documents/guidance	Yes (but document/guidance under construction)	Yes (but document/guidance under construction)	Yes (but document/guidance under construction)	De-identification followed by anonymisation, under controlled access
47						
48	3 to 5	None				De-identification followed by anonymisation, under open access
49						
50	0 to 2	Only externally sourced documents/guidance				Only de-identification, under controlled access
51	0 to 2	Both internal and external documents/guidance	Yes (documents/guidance implemented)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access
52	More than 10	Both internal and external documents/guidance	No (this process is not covered)	Yes (documents/guidance implemented)	No (this process is not covered)	De-identification followed by anonymisation, under controlled access

ID	Q12 experience in the process	Q13 experience in the process of releasing	Q14 experience in the process of maintaining released	Q15 Are you aware of or ever come across re-identification risk scores for assisting in the release	Q16 Which scenario(s) do you consider for calculating the re-identification risk score?	Q17 Please, describe how the calculated re-identification risk scores inform the process for data release.	Q18 Which program do you use for calculating the re-identification risk scores?	Q19 advantages disadvantages of using re-identification risk scores	Q20 How likely are you going to continue to use/generate the re-identification risk sco
1	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
2	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
3	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
4	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
5									
6									
7									
8	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
9	***redacted***	***redacted***	***redacted***	I have never heard of them					
10	***redacted***	***redacted***	***redacted***	I have never heard of them					
11									
12	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
13	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
14									
15	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
16	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
17	***redacted***	***redacted***	***redacted***	I have never heard of them					
18									
19									
20	***redacted***	***redacted***	***redacted***	I have never heard of them					
21	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
22	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
23	***redacted***	***redacted***	***redacted***	I have never heard of them					
24	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
25	***redacted***	***redacted***	***redacted***	I have never heard of them					
26									
27	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
28									
29	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
30	***redacted***	***redacted***	***redacted***	I have never heard of them					
31	***redacted***	***redacted***	***redacted***	I have never heard of them					
32									
33	***redacted***	***redacted***	***redacted***	I have never heard of them					
34									
35	***redacted***	***redacted***	***redacted***	I have never heard of them					
36									
37	***redacted***	***redacted***	***redacted***	I have never heard of them					
38	***redacted***	***redacted***	***redacted***	I have never heard of them					
39	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
40	***redacted***	***redacted***	***redacted***	I have never heard of them					
41	***redacted***	***redacted***	***redacted***	I have never heard of them					
42	***redacted***	***redacted***	***redacted***	I have never heard of them					
43	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
44	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
45	***redacted***	***redacted***	***redacted***	I have heard about them, but I am not so sure what they are					
46	***redacted***	***redacted***	***redacted***	I have never heard of them					
47									
48	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
49									
50	***redacted***	***redacted***	***redacted***	I have never heard of them					
51	***redacted***	***redacted***	***redacted***	I have a general understanding, but do not use them					
52	***redacted***	***redacted***	***redacted***	I have never heard of them					

ID	Q21 What are the barrier for not using the re-identification risk scores (please tick all that apply)	Q23A Data format: number of participants, data standard/software used in datasets	Q23B Data Uniqueness: are there any unique values? Is there too much granularity in the datasets?	Q23C Sensitivity: the data still contains sensitivity informatic	Q23A Motivation: Existence of a motivated individual(s)	Q23B Auxiliary information: Existence of matching datasets	Q23C Geographical location	Q23D Consequence harmful to individual	Q23K Consequence harmful to your organisation	Q24 any other aspects you consider before release
1		Always	Always	Always	Often	Rarely	Sometimes	Always	Always	***redacted***
2	Lack of funding;Lack of time ;Lack of relevant training; ***redacted***	Always	Rarely	Often	Always	Rarely	Often	Sometimes	Sometimes	***redacted***
3	Lack of time ;Lack of relevant training;	Always	Always	Always	Rarely	Rarely	Often	Always	Sometimes	***redacted***
4		Always	Always	Always	Rarely	Rarely	Always	Often	Often	***redacted***
5										
6										
7										
8		Always	Always	Always	Often	Often	Sometimes	Rarely	Sometimes	***redacted***
9		Often	Rarely	Often	Rarely	Never	Never	Sometimes	Never	
10		Always	Rarely	Always	Never	Rarely	Never	Sometimes	Never	***redacted***
11										
12	Lack of relevant training;Lack of time ;	Always	Always	Always	Rarely	Sometimes	Sometimes	Always	Often	***redacted***
13	Lack of funding;Lack of relevant training;Lack of time ; ***redacted***	Always	Often	Often	Always	Often	Rarely	Always	Always	***redacted***
14										
15		Always	Always	Always	Often	Sometimes	Sometimes	Sometimes	Sometimes	***redacted***
16		Always	Often	Always	Sometimes	Often	Rarely	Often	Sometimes	***redacted***
17		Always	Often	Always	Always	Often	Rarely	Often	Always	
18										
19										
20		Always	Always	Always	Always	Always	Sometimes	Always	Always	***redacted***
21		Often	Sometimes	Often	Rarely	Sometimes	Rarely	Often	Sometimes	***redacted***
22	Lack of funding;Lack of relevant training;Lack of time ; ***redacted***	Always	Always	Always	Always	Often	Often	Always	Always	***redacted***
23		Always	Rarely	Always	Rarely	Rarely	Never	Always	Always	***redacted***
24	Lack of relevant training;	Always	Often	Always	Always	Always	Always	Always	Always	***redacted***
25		Often	Often	Often	Often	Often	Rarely	Sometimes	Often	***redacted***
26										
27	Lack of relevant training;	Always	Often	Always	Always	Always	Rarely	Sometimes	Sometimes	***redacted***
28										
29	Lack of relevant training;Lack of time ;	Always	Rarely	Always	Sometimes	Often	Always	Always	Often	***redacted***
30		Always	Always	Always	Always	Always	Never	Always	Always	***redacted***
31		Always	Sometimes	Always	Often	Rarely	Rarely	Sometimes	Sometimes	***redacted***
32										
33		Often	Often	Always	Often	Sometimes	Often	Always	Always	***redacted***
34										
35		Always	Always	Always	Always	Always	Often	Often	Often	***redacted***
36										
37		Always	Always	Always	Sometimes	Never	Never	Sometimes	Sometimes	***redacted***
38		Often	Always	Always	Often	Sometimes	Rarely	Always	Always	***redacted***
39		Always	Rarely	Often	Sometimes	Sometimes	Sometimes	Sometimes	Sometimes	***redacted***
40		Always	Always	Always	Always	Always	Sometimes	Always	Always	***redacted***
41		Always	Always	Always	Always	Always	Often	Always	Always	
42		Always	Sometimes	Always	Rarely	Rarely	Rarely	Often	Rarely	***redacted***
43	Lack of relevant training;	Always	Always	Always	Sometimes	Rarely	Rarely	Sometimes	Always	***redacted***
44		Always	Always	Always	Sometimes	Sometimes	Rarely	Often	Often	***redacted***
45		Often	Sometimes	Sometimes	Sometimes	Sometimes	Rarely	Rarely	Sometimes	***redacted***
46		Always	Sometimes	Rarely	Always	Rarely	Rarely	Rarely	Always	***redacted***
47										
48	Lack of funding;Lack of relevant training;Lack of time ;	Rarely	Always	Always	Rarely	Sometimes	Rarely	Often	Often	***redacted***
49										
50		Always	Always	Always	Rarely	Rarely	Never	Sometimes	Sometimes	***redacted***
51	Lack of funding;Lack of relevant training;Lack of time; ***redacted***	Always	Always	Always	Sometimes	Sometimes	Often	Often	Often	***redacted***
52		Always	Always	Always	Always	Never	Never	Sometimes	Often	***redacted***

Appendix 8 - Case Study - Professor Sweeney's Research

Sweeney (Sweeney 2002) positively linked the Group Insurance Commission anonymised data (freely available to researchers) with the voter registration list for Cambridge, Massachusetts (United States (US)), which she acquired for 20 US dollars in 1996. The local press provided the final piece of information when the then Massachusetts Governor William Weld collapsed in a public appearance on the 19th May 1996, an event that was widely reported. Consequently, she knew that Governor Weld lived in Cambridge Massachusetts and according to the Cambridge voter registration list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code. This re-identified Governor Weld's anonymised medical record.

This aspect of Sweeney's research shaped US policy and due to her successful re-identification of Governor Weld's, the Health Insurance Portability and Accountability Act (HIPAA) was enacted in 1996.

This prompted Sweeney to investigate further. She demonstrated that 87% of the US population could be uniquely identified by their five-digit ZIP code, sex and date of birth. So, even though each of these variables on their own would be non-identifiable, storing/linking them together makes it possible to uniquely identify an individual (Wes 2017). Clearly, data released containing such information about these individuals should not be automatically considered anonymous. This concept of uniqueness could also apply to the UK population (or in the worst case, potentially underestimate it due to the smaller population size when compared to the US), as the UK is organised in a similar manner with postcodes.

Sweeney has continued her research, and more than 15 years after the introduction of HIPAA, she shown in 2013 that improvements are still needed to achieve anonymity in datasets, as she managed to re-identify a patient in the Washington State's health records using linkage to traditional data sources (newspaper articles and public records) (Sweeney 2013). In 2017, she indicated that HIPAA "is not sufficient to protect data against re-identification" (Sweeney, Yoo et al. 2017).

Sweeney's team's latest project theDataMap™ (Sweeney, Zang et al. 2010), provides a sobering overview of all the places US patients' data goes, sometimes whether they want it to or not, given the advancement of the internet. Most of the paths travelled by the data are relatively innocuous, but there are interactions that can be harmful to the individual, for example, employers or financial institutions may get access to personal health information that individuals would prefer to be keep away from such organisations, as it could affect their employability of credit worthiness.

Reference list

Agencia Española Proteccion Datos and The European Data Protection Supervisor (EDPS) (2021). 10 MISUNDERSTANDINGS RELATED TO ANONYMISATION.
Al-Shahi, R. (2000). "Using patient-identifiable data for observational research and audit." BMJ **321**: 1031–1032.
Anbazhagan, K., R. Sugumar, M. Mahendran and R. Natarajan (2012). "An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining." International Journal of Advanced Research in Computer and Communication Engineering(7).
Anderson, M. L., K. Chiswell, E. D. Peterson, A. Tasneem, J. Topping and R. M. Califf (2015). "Compliance with results reporting at ClinicalTrials. gov." New England Journal of Medicine **372**(11): 1031-1039.
Aryanto, K. Y. E., G. van Kernebeek, B. Berendsen, M. Oudkerk and P. M. A. van Ooijen (2016). "Image De-Identification Methods for Clinical Research in the XDS Environment." Journal of Medical Systems **40**(4).
Azizi, Z., C. Zheng, L. Mosquera, L. Pilote and K. El Emam (2021). "Can synthetic data be a proxy for real clinical trial data? A validation study." BMJ open **11**(4): e043497.
Babbie, E. R. (2020). Chapter 4 - Research Design. The practice of social research, Cengage**:** 575.
Bagley, H. J., H. Short, N. L. Harman, H. R. Hickey, C. L. Gamble, K. Woolfall, B. Young and P. R. Williamson (2016). "A patient and public involvement (PPI) toolkit for meaningful and flexible involvement in clinical trials–a work in progress." Research involvement and engagement **2**: 1-14.
Balas, E. A., M. Vernon, F. Magrabi, L. T. Gordon and J. Sexton (2015). Big Data Clinical Research: Validity, Ethics, and Regulation. Medinfo 2015: Ehealth-Enabled Health. I. N. Sarkar, A. Georgiou and P. M. D. Marques. **216:** 448-452.
Bamford, S., S. Lyons, L. Arbuckle and P. Chetelat (2022). "Sharing Anonymized and Functionally Effective (SAFE) data standard for safely sharing rich clinical trial data."
BBC News (2020). NHS data breach involving 284 patients uncovered. BBC News. Online UK.
Bertagnolli, M., O. Sartor, B. Chabner, M. Rothenberg, S. Khozin, C. Hugh-Jones, D. M. Reese and M. J. Murphy (2017). "Advantages of a truly open-access data-sharing model." N Engl J Med **12**(376): 1178-1181.
Bhandari, P. (2022). "What Is Quantitative Research? | Definition & Methods." 2024, from https://www.scribbr.co.uk/research-methods/introduction-to-quantitative-research/.
Bhatt, A. (2010). "Evolution of clinical research: a history before and beyond James Lind." Perspectives in clinical research **1**(1): 6-10.
Bin, Z., P. Jian and L. Wo-Shun (2008). "A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data." ACM SIGKDD Explorations **10**(2): 12-22.
BioMed Central Ltd. "ISRCTN registry." Retrieved 13 Mar 2024, 2024, from https://www.isrctn.com/.
Biomedical, U. S. N. C. f. t. P. o. H. S. o. and B. Research (1978). The Belmont report: ethical principles and guidelines for the protection of human subjects of research, Department of Health, Education, and Welfare, National Commission for the ….
Black, N. (2013). "Patient reported outcome measures could help transform healthcare." Bmj **346**.
Bothwell, L. E., J. A. Greene, S. H. Podolsky and D. S. Jones (2016). Assessing the gold standard—lessons from the history of RCTs, The New England Journal of Medicine,. **374:** 2175-2181.

Brody, T. (2016). Clinical trials: study design, endpoints and biomarkers, drug safety, and FDA and ICH guidelines, Academic press.
Bruckner, T. (2017). "Clinical trial transparency: a guide for policy makers." UK: Transparency International Pharmaceuticals & Healthcare Programme.
Butte, A. J. (2021). "Trials and tribulations-11 reasons why we need to promote clinical trials data sharing." JAMA Network Open **4**(1): e2035043-e2035043.
Califf, R. M. (2023). "The Importance of Clinical Trial Transparency and FDA Oversight." Retrieved 14 Mar 2024, 2024, from https://www.fda.gov/news-events/fda-voices/importance-clinical-trial-transparency-and-fda-oversight.
Cancer Institute New South Wales (NSW). (2022). "How do clinical trials progress?" Retrieved 30 Oct 2022, 2022, from https://www.cancer.nsw.gov.au/research-and-data/cancer-clinical-trials-in-nsw/how-do-clinical-trials-progress.
Cancer Research UK. "Data sharing guidelines." Retrieved 30 Oct 2020, 2020, from https://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy/data-sharing-guidelines.
Cancer Research UK. (2016). "Why aren't we sharing? ." 2018, from https://www.cancerresearchuk.org/funding-for-researchers/research-features/2016-08-10-why-arent-we-sharing.
Cancer Research UK. (2022). "What are clinical trials?" Retrieved 14 March 2024, 2024, from https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/what-clinical-trials-are.
Chaturvedi, N., B. Mehrotra, S. Kumari, S. Gupta, H. Subramanya and G. Saberwal (2019). "Some data quality issues at ClinicalTrials. gov." Trials **20**: 1-8.
Chevrier, R., V. Foufi, C. Gaudet-Blavignac, A. Robert and C. Lovis (2019). "Use and understanding of anonymization and de-identification in the biomedical literature: scoping review." Journal of medical Internet research **21**(5): e13484.
Cingi, C. and N. Bayar Muluk (2016). Quick Guide to Good Clinical Practice, Springer.
Clinical Data Interchange Standards Consortium. "Analysis Data Model (ADaM)." Retrieved 22 Mar 2024, 2024, from https://www.cdisc.org/standards/foundational/adam.
Clinical Data Interchange Standards Consortium. "CDISC Foundational Standards." Retrieved 22 Mar 2024, 2024, from https://www.cdisc.org/standards/foundational.
Clinical Data Interchange Standards Consortium. "Controlled Terminology." Retrieved 22 Mar 2024, 2024, from https://www.cdisc.org/standards/terminology/controlled-terminology.
Clinical Data Interchange Standards Consortium. "Study Data Tabulation Model (SDTM)." Retrieved 22 Mar 2024, 2024, from https://www.cdisc.org/standards/foundational/sdtm.
Clinical Trials a SAGE journal. (2022). "Metrics and citations for: Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review." Retrieved 08 May 2024, 2024, from https://journals.sagepub.com/doi/10.1177/17407745221087469#core-collateral-metrics.
Creswell, J. W. (2008). "Planning, conducting and evaluating quantitative and quantitative research." Educational Research. Upper Saddle River, NJ: Pearson Education Inc.
Custers, B., A. M. Sears, F. Dechesne, I. Georgieva, T. Tani and S. Van der Hof (2019). EU personal data protection in policy and practice, Springer.
Dal-Ré, R., R. Banzi, I. A. Cristea, C. Fernández-de-Las-Peñas, L. G. Hemkens, P. Janiaud, M. S. Jansen, F. Naudet and F. R. Rosendaal (2023). "Using the phases of clinical development of medicines to describe clinical trials assessing other interventions is widespread but not useful." Journal of Clinical Epidemiology **161**: 157-163.
Dal-Ré, R. and I. Mahillo-Fernández (2023). "Posting of clinical trial results and other critical information from completed medicines trials on ClinicalTrials. gov." European Journal of Clinical Pharmacology **79**(10): 1385-1390.
Data Protection Commission - Ireland. "Anonymisation and pseudonymisation " Retrieved 20 JUN 2018, 2018, from https://www.dataprotection.ie/docs/Anonymisation-and-pseudonymisation/1594.htm.
Davies, R. (2023). Capita cyber-attack: USS pension fund members' details may have been stole. The Guardian. Online UK.

Digital Curation Centre. "Research Funding Policies_Medical Research Council (MRC)." Retrieved 08 JUL 2018, 2018, from http://www.dcc.ac.uk/resources/policy-and-legal/research-funding-policies/mrc.
Digital Curation Centre. (2023). "Overview of funders' data policies." Retrieved 23 MAR 2024, 2024, from https://www.dcc.ac.uk/guidance/policy/overview-funders-data-policies.
Doel, T., D. I. Shakir, R. Pratt, M. Aertsen, J. Moggridge, E. Bellon, A. L. David, J. Deprest, T. Vercauteren and S. Ourselin (2017). "GIFT-Cloud: A data sharing and collaboration platform for medical imaging research." COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE **139**: 181-190.
Duckworth, A. D., M. M. McQueen, C. E. Tuck, J. H. Tobias, J. M. Wilkinson, L. C. Biant, E. C. Pulford, S. Aldridge, C. Edwards and C. P. Roberts (2019). "Effect of alendronic acid on fracture healing: a multicenter randomized placebo-controlled trial." Journal of Bone and Mineral Research **34**(6): 1025-1032.
Dwork, C. (2011). Differential Privacy. Encyclopedia of Cryptography and Security, Springer, Boston, MA**:** 338-340.
Eichler, H.-G. and G. Rasi (2020). "Clinical trial publications: a sufficient basis for healthcare decisions?" European Journal of Internal Medicine **71**: 13-14.
El Emam, K. (2011). "Methods for the de-identification of electronic health records for genomic research." Genome Medicine **3**(4): 25.
El Emam, K. (2013). Chapter 16 - Measuring the Probability of Re-Identification. Guide to the de-identification of personal health information, CRC Press.
El Emam, K. (2013). Guide to the de-identification of personal health information, CRC Press.
El Emam, K. and K. Abdallah (2015). "De-identifying data in clinical trials." Applied Clinical Trials **24**(8/9): 40.
El Emam, K. and L. Arbuckle (2013). Anonymizing health data : case studies and methods to get you started, O'Reilly Media, Inc.
El Emam, K., L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri, S. Rose, J. Howard and J. Gluck (2012). "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset." J Med Internet Res **14**(1): e33.
El Emam, K., D. Buckeridge, R. Tamblyn, A. Neisa, E. Jonker and A. Verma (2011). "The re-identification risk of Canadians from longitudinal demographics." BMC Medical Informatics and Decision Making **11**(1): 46.
El Emam, K. and A. Fineberg (2009). "An Overview of Techniques for De-identifying Personal Health Information." CHEO Research Institute.
El Emam, K., M. Hintze and R. Boardman (2020). "Does de-identification require consent under the GDPR and English common law?" Journal of Data Protection & Privacy **3**(3): 291-298.
El Emam, K., S. Jabbouri, S. Sams, Y. Drouet and M. Power (2006). "Evaluating Common De-Identification Heuristics for Personal Health Information." Journal of Medical Internet Research **8**(4): 4-4.
El Emam, K., E. Jonker, L. Arbuckle and B. Malin (2011). "A Systematic Review of Re-Identification Attacks on Health Data." Plos one.
El Emam, K., L. Mosquera and C. Zheng (2020). "Optimizing the synthesis of clinical trial data using sequential trees." Journal of the American Medical Informatics Association.
El Emam, K., S. Rodgers and B. Malin (2015). "Anonymising and sharing individual patient data." BMJ **350**: h1139.
Elliot, M., E. Mackey and K. O'Hara (2020). The Anonymisation Decision-making Framework: European practitioners' guide.
Elliot, M. M., E; O'Hara, K.; Tudor-Smith C. (2016). The Anonymisation Decision-Making Framework (ADF). UK, 1st edUKAN Publications.
Erlich, Y. (2013). "Breaking Good: A Short Ethical Manifesto for the Privacy Researcher." Bill of Health's symposium on the Law, Ethics, and Science of Re-Identification Demonstrations http://blogs.harvard.edu/billofhealth/2013/05/23/breaking-good-a-short-ethical-manifesto-for-the-privacy-researcher/ 2018.

Esmail, L. C., P. Kapp, R. Assi, J. Wood, G. Regan, P. Ravaud and I. Boutron (2023). "Sharing of individual patient-level data by trialists of randomized clinical trials of pharmacological treatments for COVID-19." JAMA **329**(19): 1695-1697.
European Commission (2014). Opinion 05/2014 on Anonymisation Techniques- ARTICLE 29 DATA PROTECTION WORKING PARTY.
European Medicines Agency (EMA) (2019). European Medicines Agency policy on publication of clinical data for medicinal products for human use POLICY/0070. EMA/144064/2019**:** 22.
European Parliament - Council of the European Union (2016). REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union**:** 1-78.
Evans, B. J. and G. P. Jarvik (2018). "Impact of HIPAA's minimum necessary standard on genomic data sharing." Genetics in Medicine **20**(5): 531-535.
Evans, S. R. (2010). "Fundamentals of clinical trial design." Journal of experimental stroke & translational medicine **3**(1): 19.
Ferran, J.-M. and S. J. Nevitt (2019). "European medicines Agency policy 0070: an exploratory review of data utility in clinical study reports for academic research." BMC medical research methodology **19**: 1-10.
Finck, M. and F. Pallas (2020). "They who must not be identified—distinguishing personal from non-personal data under the GDPR." International Data Privacy Law **10**(1): 11-36.
Fujimoto, H. (2023). "The Crucial Role of Clinical Trials in Advancing Medical Knowledge and Treatment." International Research Journal of Pharmacy and Pharmacology **11**(3): 3.
Gabelica, M., R. Bojčić and L. Puljak (2022). "Many researchers were not compliant with their published data sharing statement: a mixed-methods study." Journal of Clinical Epidemiology **150**: 33-41.
Gallagher, J. (2023). "New superbug-killing antibiotic discovered using AI." Retrieved 23 APR 2024, 2024, from https://www.bbc.co.uk/news/health-65709834.
Gamertsfelder, E., N. D. Figueroa, S. Keestra, A. R. Silva, R. Borana, M. Siebert and T. Bruckner (2023). "Towards transparency: adoption of WHO best practices in clinical trial registration and reporting among top medical research funders in the USA." BMJ evidence-based medicine.
Garfinkel, S. L. (2015). De-Identification of Personal Information. N. I. o. S. a. T. I. Report. US, US Department of Commerce**:** 46.
George, T. (2021). "Exploratory Research | Definition, Guide, & Examples." 2024, from https://www.scribbr.com/methodology/exploratory-research/.
Goldacre, B., N. J. DeVito, C. Heneghan, F. Irving, S. Bacon, J. Fleminger and H. Curtis (2018). "Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource." bmj **362**.
Goldacre, B., S. Lane, K. R. Mahtani, C. Heneghan, I. Onakpoya, I. Bushfield and L. Smeeth (2017). "Pharmaceutical companies' policies on access to trial data, results, and methods: audit study." bmj **358**.
Goodwin, J. (2012). SAGE Secondary Data Analysis. London.
Google. "Google schoolar search results for doi 10.1177/17407745221087469." Retrieved 08 May 2024, 2024, from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=10.1177%2F17407745221087469&btnG=.
Google. (2024). "What is Big Data?" Retrieved 27 May 2024, 2024, from https://cloud.google.com/learn/what-is-big-data.
Gøtzsche, P. C. (2011). "Why we need easy access to all data." Trials.
Government of the Netherlands. "What is Brexit?" Retrieved 30 Mar 2024, 2024, from https://www.government.nl/topics/european-union/question-and-answer/what-is-brexit.
Gretel Labs. (2024). "What is Synthetic Data?" Retrieved 24 Mar 2024, 2024, from https://gretel.ai/what-is/synthetic-data.

Gudi, N., P. Kamath, T. Chakraborty, A. G. Jacob, S. S. Parsekar, S. N. Sarbadhikari and O. John (2022). "Regulatory frameworks for clinical trial data sharing: scoping review." Journal of medical Internet research **24**(5): e33591.
Han, J., J. Yu, J. Lu, H. Peng and J. Wu (2017). "An Anonymization Method to Improve Data Utility for Classification." **10581**: 57-71.
Handelsman, D. (2015). "Now You See It, Now You Don't - Using SAS to De-Identify Data to Support Clinical Trial Data Transparency."
Hanson, H. (2018). "Clinical trial results disclosure on ClinicalTrials. gov and EudraCT." Medical Writing **27**: 44-48.
Hartung, D. M., D. A. Zarin, J.-M. Guise, M. McDonagh, R. Paynter and M. Helfand (2014). "Reporting discrepancies between the ClinicalTrials. gov results database and peer-reviewed publications." Annals of internal medicine **160**(7): 477-483.
Health, U. D. o. and H. Services (2003). "Protecting personal health information in research: Understanding the HIPAA Privacy Rule." Washington, DC: Author.
Healy, C. (2019). University ordered to disclose children's clinical trial data. eAlerts. C. Law-Now™. Online UK.
HMA Permanent Secretariat (2021). Subject: Proposal to the Heads of Medicines Agencies to improve harmonisation of access to Clinical Study Reports across National Competent Authorities.
Holtedahl, R. (2020). "Failure to report results of registered trials of treatments for shoulder complaints." Tidsskrift for Den norske legeforening.
Howe, N. (2021). Participants' Attitudes Towards Data Sharing. Doctor of Philosophy, Newcastle University.
Howe, N., E. Giles, D. Newbury-Birch and E. McColl (2018). "Systematic review of participants' attitudes towards data sharing: a thematic synthesis." Journal of health services research & policy **23**(2): 123-133.
Hrynaszkiewicz, I., M. L. Norton, A. J. Vickers and D. G. Altman (2010). "Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers." Trials **11**(340).
Hrynaszkiewicz, I., N. Simons, A. Hussain, R. Grant and S. Goudie (2020). "Developing a research data policy framework for all journals and publishers." Data Science Journal **19**: 5-5.
Hughes, S., K. Wells, P. McSorley and A. Freeman (2014). "Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach." Pharmaceutical Statistics **13**(3): 179-183.
Humphreys, G. S., G. Merriott , R. Knowles, B. Pierson and P. Quattroni (2020). Clinical Trial Data Sharing: What We've Heard from Researchers. figshare. Report.
Huser, V. and D. Shmueli-Blumberg (2018). "Data sharing platforms for de-identified data from human clinical trials." Clinical Trials **15**(4): 413-423.
Information Commissioner's Office (ICO). "Who we are." Retrieved 08 JUL 2018, 2018, from https://ico.org.uk/about-the-ico/who-we-are/.
Information Commissioner's Office (ICO) (2012). Anonymisation: managing data protection risk-code of practice. UK.
Information Commissioner's Office (ICO) (2021). Anonymisation, pseudonymisation and privacy enhancing technologies guidance - Draft. UK.
Information Commissioner's Office (ICO) (2022). Data sharing - code of practice.
Information Commissioner's Office (ICO) (2023). A guide to lawful basis. UK.
Information Commissioner's Office (ICO) (2023). A guide to the data protection principles. UK. **2023**.
information Commissioner's Office (ICO). (2023). "Special category data." Retrieved 01 Mar 2024, 2024.
International Clinical Trials Methodology Conference (ICTCM), M. Sydes, K. Gillies and P. Williamson (2023). "ICTMC 2022: 6th International Clinical Trials Methodology Conference - book of abstracts - PS8A-01 - What are the re-identification risk scores of publicly available anonymised clinical trial datasets? ." ZENODO ( ): .

International Committee of Medical Journal Editors (ICMJE) (2024). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Updated January 2024.
IOM (Institute of Medicine) (2015). Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington (DC), National Academies Press.
Jiang, Y., L. Mosquera, B. Jiang, L. Kong and K. El Emam (2022). "Measuring re-identification risk using a synthetic estimator to enable data sharing." PLOS ONE **17**(6): e0269097.
Kaplan, R. M., A. J. Koong and V. Irvin (2023). "Food and Drug Administration novel drug decisions in 2017: transparency and disclosure prior to and 5 years following approval." Health Affairs Scholar **1**(2).
Keerie, C., C. Tuck, G. Milne, S. Eldridge, N. Wright and S. C. Lewis (2018). "Data sharing in clinical trials - practical guidance on anonymising trial datasets." Trials **19**(1): 25.
Khan, S. (2023). NHS secretary fined for accessing Worcestershire medical records. BBC News. Online UK.
Kulakiewicz, A., E. Parkin and T. Powell (2022). "Patient health records: Access, sharing and confidentiality." House of Commons Research Briefing(07103).
Law, M., D.-L. Couturier, B. Choodari-Oskooei, P. Crout, C. Gamble, P. Jacko, P. Pallmann, M. Pilling, D. S. Robertson and M. Robling (2023). "Medicines and Healthcare products Regulatory Agency's "Consultation on proposals for legislative changes for clinical trials": a response from the Trials Methodology Research Partnership Adaptive Designs Working Group, with a focus on data sharing." Trials **24**(1): 640.
Law, M., D. Couturier, B. Choodari-Oskooei, P. Crout, C. Gamble, P. Pallmann, M. Pilling, D. Robertson, M. Robling and M. R. Sydes (2022). "Response to Medicines and Healthcare products Regulatory Agency's "Consultation on proposals for legislative changes for clinical trials"."
Li, N., T. Li and S. Venkatasubramanian (2007). "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity." 2007 IEEE 23rd International Conference on Data Engineering, Istanbul **2**: 106-115.
Loder, E., H. Macdonald, T. Bloom and K. Abbasi (2024). Mandatory data and code sharing for research published by The BMJ, British Medical Journal Publishing Group. **384**.
Longo, D. L. and J. M. Drazen (2016). "Data Sharing." N Engl J Med **374**(3): 276-277.
Machanavajjhala, A., D. Kifer, J. Gehrke and M. Venkitasubramaniam (2007). "l-diversity; Privacy Beyond k-Anonymity." ACM Transactions on Knowledge Discovery from Data **1**(1): 3-es.
Malin, B. A., K. El Emam and C. M. O'Keefe (2013). "Biomedical data privacy: problems, perspectives, and recent advances." Journal of the American Medical Informatics Association : JAMIA **20**(1): 2-6.
Mann, G., T. J. Pedersen, R. Lyzinski, A. Scott, A. J. Foglia, J. Cromer, M. Dong, N. Varga, S. Gardner and C. J. Kirchberg (2023). "CDISC Enables Efficient Streamlining of Clinical Trial Safety Evaluation." Journal of the Society for Clinical Data Management **3**(1).
McCarthy, M., K. Gillies, N. Rousseau, J. Wade, C. Gamble, E. Toomey, K. Matvienko-Sikar, M. Sydes, M. Dowling, V. Bryant, L. Biesty and C. Houghton (2023). "Qualitative data sharing practices in clinical trials in the UK and Ireland: towards the production of good practice guidance." HRB Open Res **6**: 10.
Medical Research Council (MRC) (2023). MRC data sharing policy.
Medicines and Healthcare products Regulatory Agency (MHRA) (2023). Government response to Consultation on proposals for legislative changes for clinical trials.
Moberg, J. and M. Kramer (2015). "A brief history of the cluster randomised trial design." Journal of the Royal Society of Medicine **108**(5): 192-198.
Modi, N. D., G. Kichenadasse, T. C. Hoffmann, M. Haseloff, J. M. Logan, A. A. Veroniki, R. L. Venchiarutti, A. K. Smit, H. Tuffaha and H. Jayasekara (2023). "A 10-year update to the principles for clinical trial data sharing by pharmaceutical companies: perspectives based on a decade of literature and policies." BMC medicine **21**(1): 400.

MRC Clinical Trials Unit - UCL. (2024). "What is a randomised clinical trial?" Retrieved 24 Mar 2024, 2024, from https://www.mrcctu.ucl.ac.uk/patients-public/about-clinical-trials/what-is-a-randomised-clinical-trial/.
Murad, M. H., N. Asi, M. Alsawas and F. Alahdab (2016). "New evidence pyramid." BMJ Evidence-Based Medicine **21**(4): 125-127.
Narayanan, A. and E. W. Felten (2014). "No silver bullet: De-identification still doesn't work." White Paper: 1-8.
National Committee on Vital and Health Statistics (2017). Recommendations on De-identification of Protected Health Information under HIPAA. D. o. H. a. H. Services.
National Institute for Health and Care Research (NIHR). (2019). "Clinical Trials Guide." Retrieved 12 MAR 2024, 2024, from https://www.nihr.ac.uk/documents/clinical-trials-guide/20595.
National Institute for Health and Care Research (NIHR). (2019). "NIHR position on the sharing of research data." Retrieved 23 Mar 2024, 2024, from https://www.nihr.ac.uk/documents/nihr-position-on-the-sharing-of-research-data/12253.
National Institute of Health (NIH) - Department of Health and Human Services (HHS) (2016). 42 CFR Part 11, Clinical Trials Registration and Results Information Submission; Final Rule. Vol. 81, No. 183. US, Federal Register**:** 64982–65157.
National Institutes of Health (NIH) (2023). Final NIH Policy for Data Management and Sharing. N. S. D. Sharing. USA.
NHS. "Clinical trials." Retrieved 13 March 2024, 2024, from https://www.nhs.uk/conditions/clinical-trials/.
NIH U.S. National Library of Medicine. (2024). "About ClinicalTrials.gov." Retrieved 19 Mar 2024, 2024, from https://clinicaltrials.gov/about-site/about-ctg.
O'Keefe, C. M., S. Otorepec, M. Elliot, E. Mackey and K. O'Hara (2017). The de-identification decision-making framework, CSIRO Reports EP173122 and EP175702 Canberra: Australian Government Office ….
Office of the Information Commissioner Queensland (2020). Privacy and Public Data - Managing re-identification risk. Queensland, Australia.
Ohmann, C., R. Banzi, S. Canham, S. Battaglia, M. Matei, C. Ariyo, L. Becnel, B. Bierer, S. Bowers, L. Clivio, M. Dias, C. Druml, H. Faure, M. Fenner, J. Galvez, D. Ghersi, C. Gluud, T. Groves, P. Houston, G. Karam, D. Kalra, R. L. Knowles, K. Krleža-Jerić, C. Kubiak, W. Kuchinke, R. Kush, A. Lukkarinen, P. S. Marques, A. Newbigging, J. O'Callaghan, P. Ravaud, I. Schlünder, D. Shanahan, H. Sitter, D. Spalding, C. Tudur-Smith, P. van Reusel, E.-B. van Veen, G. R. Visser, J. Wilson and J. Demotes-Mainard (2017). "Sharing and reuse of individual participant data from clinical trials: principles and recommendations." BMJ Open **7**(12).
Packer, M. (2016). "Data sharing: lessons from Copernicus and Kepler." BMJ **354**: i4911.
Palmer, S. (2024) "Empowering patients through more accessible clinical trial information."
Personal Data Protection Commission Singapore (2018). Guide to basic data anonymisation techniques.
Peters, M. D., C. M. Godfrey, H. Khalil, P. McInerney, D. Parker and C. B. Soares (2015). "Guidance for conducting systematic scoping reviews." JBI Evidence Implementation **13**(3): 141-146.
Peterson, J., P. F. Pearce, L. A. Ferguson and C. A. Langford (2017). "Understanding scoping reviews: Definition, purpose, and process." Journal of the American Association of Nurse Practitioners **29**(1): 12-16.
PHUSE, P., Kayley (2020). "The Impacts of COVID-19 on Clinical Trial Transparency and Document Disclosure PHUSE CTT Project." Retrieved 13 JAN 2021, 2021, from https://phuse.global/Communications/PHUSE_Blog/the-impacts-of-covid-19-on-clinical-trial-transparency-and-document-disclosure-phuse-ctt-project.
Pilgram, L., T. Meurers, B. Malin, GCKD Investigators, E. Schaeffner, K. Eckardt and F. Prasser (2024). PRESENTATION: The Costs of Anonymization: Case Sutdy Using Clinical Data. J Med Internet Res (forthcoming), CHEO Research Institute and the University of Ottawa.

Pilgram, L., T. Meurers, B. Malin, E. Schaeffner, K.-U. Eckardt, F. Prasser and G. Investigators (2024). "The Costs of Anonymization: Case Study Using Clinical Data." Journal of Medical Internet Research **26**: e49445.
Pisani, E., P. Aaby, J. G. Breugelmans, D. Carr, T. Groves, M. Helinski, D. Kamuya, S. Kern, K. Littler, V. Marsh, S. Mboup, L. Merson, O. Sankoh, M. Serafini, M. Schneider, V. Schoenenberger and P. J. Guerin (2016). "Beyond open data: realising the health benefits of sharing data." BMJ **355**: i5295.
Pisternick-Ruf, W., A. Marquart and T. M. Schindler (2018). "Clinical data publication by the EMA: The challenges facing the pharmaceutical industry." Medical Writing **27**: 39-43.
Potiguara Carvalho, A., F. Potiguara Carvalho, E. Dias Canedo and P. H. Potiguara Carvalho (2020). Big data, anonymisation and governance to personal data protection. The 21st Annual International Conference on Digital Government Research.
Prasser, F., R. Bild and K. A. Kuhn (2016). A Generic Method for Assessing the Quality of De-Identified Health Data. MIE.
Precedence Research Pvt Ltd. (2024). "Clinical Trials Market Size, Growth, Trends, Report By 2032." Retrieved 17 Mar 2024, 2024, from https://www.precedenceresearch.com/clinical-trials-market.
PrivazyPlan®. (2018). "Recital 26 EU General Data Protection Regulation (EU-GDPR). Privacy_Privazy according to plan." Retrieved 08 JUL 2018, 2018, from http://www.privacy-regulation.eu/en/recital-26-GDPR.htm.
Raghunathan, B. (2013). The complete book of data anonymization: from planning to implementation, CRC Press.
Rios, R. S., K. I. Zheng and M.-H. Zheng (2020). "Data sharing during COVID-19 pandemic: what to take away." Expert Review of Gastroenterology & Hepatology **14**(12): 1125-1130.
Rodriguez, A., S. C. Lewis, S. Eldridge, T. Jackson and C. J. Weir (2024). "A survey on UK researchers' views regarding their experiences with the de-identification, anonymisation, release methods and re-identification risk estimation for clinical trial datasets." Clinical Trials **0 : Ahead of print**(0): 13.
Rodriguez, A., C. Tuck, M. F. Dozier, S. C. Lewis, S. Eldridge, T. Jackson, A. Murray and C. J. Weir (2022). "Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review." Clinical Trials **19**(4): 452-463.
Ross, J. S., T. Tse, D. A. Zarin, H. Xu, L. Zhou and H. M. Krumholz (2012). "Publication of NIH funded trials registered in ClinicalTrials. gov: cross sectional analysis." Bmj **344**.
Rowhani-Farid, A., M. Grewal, S. Solar, A. O. Eghrari, A. D. Zhang, C. P. Gross, H. M. Krumholz and J. S. Ross (2023). "Clinical trial data sharing: a cross-sectional study of outcomes associated with two US National Institutes of Health models." Scientific Data **10**(1): 529.
Sariyar, M., I. Schluender, C. Smee and S. Suhr (2015). "Sharing and reuse of sensitive data and samples: supporting researchers in identifying ethical and legal requirements." Biopreservation and biobanking **13**(4): 263-270.
SAS Institute Inc (2013). SAS 9.4 [Computer software] TS level 1M4, Copyright © 2016 SAS Institute Inc. Cary, NC, USA, SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Saunders, M., P. Lewis and A. Thornhill (2009). Chapter 8 - Utilising secondary data. Research methods for business students, Pearson education.
Scottish Longitudinal Study Development & Support Unit. (2013). "Synthetic Data Estimation for UK Longitudinal Studies " Retrieved 08 Jul 2018, 2018, from https://sls.lscs.ac.uk/projects/view/2013_012/.
Sheard, J. (2018). Chapter 18 - Quantitative data analysis. Research Methods (Second Edition). K. Williamson and G. Johanson, Chandos Publishing**:** 429-452.
Simon, G. E., S. M. Shortreed, R. Y. Coley, R. B. Penfold, R. C. Rossom, B. E. Waitzfelder, K. Sanchez and F. L. Lynch (2019). "Assessing and minimizing re-identification risk in research data derived from health care records." eGEMs **7**(1).

Skinner, C. (2009). Statistical Disclosure Control for Survey Data. Handbook of Statistics: Sample Surveys: Design, Methods and Applications. I. D. P. C. R. R. eds. Amsterdam:, Elsevier**:** 381-396.
Stalla-Bourdillon, S. and A. Knight (2016). "Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data." Wis. Int'l LJ **34**: 284.
Swedberg, R. (2020). "Exploratory research." The production of knowledge: Enhancing progress in social science **2**(1): 17-41.
Sweeney, L. (2002). "Achieving k-anonymity privacy protection using generalization and suppression."
Sweeney, L. (2002). "k-Anonymity: a model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.
Sweeney, L. (2013). "Matching Known Patients to Health Records in Washington State Data." Harvard University. Data Privacy Lab. White Paper **1089-1**.
Sweeney, L., A. Abu and J. Winn (2013). "Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment)." Harvard University. Data Privacy Lab. White Paper arXiv.org **1021-1**.
Sweeney, L., J. S. Yoo, L. Perovich, K. E. Boronow, P. Brown and J. G. Brody (2017). "Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study." Technology science **2017**.
Sweeney, L., J. Zang, J. S. Yoo and S. Hooley. (2010). "About-theDataMap." Retrieved 08 JUL 2018, 2018, from https://thedatamap.org/about.php.
Templ, M., A. Kowarik and B. Meindl (2015). "Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro." Journal of Statistical Software **67**(4): 1-36.
Templ, M., B. Meindl and A. Kowarik (2021). Package sdcMicro.
The Crown Prosecution Service. (2018). "Data Protection Act 2018 - Criminal Offences." 2023, from https://www.cps.gov.uk/legal-guidance/data-protection-act-2018-criminal-offences.
The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) (2016). Integrated addendum to ICH E6(R1): Guideline for Good Clinical Practice E6(R2). ICH HARMONISED GUIDELINE. **E6(R2)**.
The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) (2021). General Considerations for Clinical Studies E8(R1). ICH HARMONISED GUIDELINE. **E8(R1)**.
The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) (2021). ICH-E6 Good Clinical Practice (GCP) - Explanatory Note.
The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) (2023). Good Clinical Practice (GCP) E6(R3) - Draft Version. ICH HARMONISED GUIDELINE. **E6(R3)**.
The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). (2024). "Welcome to the ICH Official Website." Retrieved 17 Mar 2024, 2024, from https://www.ich.org/.
The Joanna Briggs Institute (2015). Joanna Briggs Institute Reviewers' Manual: 2015 edition / Supplement. Methodology for JBI Scoping Reviews, The Joanna Briggs Institute.
The National Archives (2016). United Kingdom General Data Protection Regulation (UK GDPR) - Regulation (EU) 2016/679 of the European Parliament and of the Council - Regulations originating from the EU.
The National Archives (2018). Data Protection Act 2018.
Trezza, D. (2023). "To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research." Frontiers in Sociology **8**: 1145038.
Tucker, K., J. Branson, M. Dilleen, S. Hollis, P. Loughlin, M. J. Nixon and Z. Williams (2016). "Protecting patient privacy when sharing patient-level data from clinical trials." BMC Medical Research Methodology **16 Suppl 1**: 77.
Tudur Smith, C., C. Hopkins, M. Sydes, K. Woolfall, M. Clarke, G. Murray and P. Williamson (2015). "Good practice principles for sharing individual participant data from publicly funded clinical trials - Oral Presentation." Trials **16**(S2).

Tudur Smith, C., C. Hopkins, M. R. Sydes, K. Woolfall, M. Clarke, G. Murray and P. Williamson (2015). Good practice principles for sharing individual participant data from publicly funded clinical trials Medical Research Council - Hubs for Trials Methodology Research.
U.S. Department of Health & Human Services (HHS) (2012). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.
U.S. Department of Health & Human Services (HHS). (2022). "The HIPAA Privacy Rule." Retrieved 13 MAR 2024, 2024, from https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.
U.S. Government (1996). Health Insurance Portability and Accountability Act of 1996. Public law. **104:** 191.
U.S. Government Printing Office (1949). Nuremberg Code. Trials of war criminals before the Nuremberg military tribunals under control council law No. 10. Washington, D.C. **2:** 181-182.
UK Clinical Research Collaboration (UKCRC). (2023). "Clinical Trials Units." Retrieved 05 Dec 2023, 2023, from https://www.ukcrc.org/research-infrastructure/clinical-trials-units/.
UKCRC. (2024). "UKCRC Registered Clinical Trials Units." Retrieved 27 May 2024, 2024, from https://www.ukcrc.org/research-infrastructure/clinical-trials-units/registered-clinical-trials-units/.
UKCRC. (2024). "What is the UKCRC?" Retrieved 27 May 2024, 2024, from https://www.ukcrc.org/about-the-ukcrc/what-is-the-ukcrc/.
van Panhuis, W., P. Paul, C. Emerson, J. Grefenstette, R. Wilder, A. Herbst, D. Heymann and D. Burke (2014). "A systematic review of barriers to data sharing in public health." BMC Public Health **14**: 1144.
Vassar, M., S. Jellison, H. Wendelbo and C. Wayant (2020). "Data sharing practices in randomized trials of addiction interventions." Addictive behaviors **102**: 106193.
Vivli Center for Global Clinical Research Data. (2020). "Vivli, a global data-sharing and analytics platform. ." 2020, from https://vivli.org/.
Vogt, W. (2011). SAGE Quantitative Research Methods. Thousand Oaks.
Vorland, C. J., A. W. Brown, H. Kilicoglu, X. Ying and E. Mayo-Wilson (2024). "Publication of Results of Registered Trials With Published Study Protocols, 2011-2022." JAMA network open **7**(1): e2350688-e2350688.
Wes, M. (2017). "Looking to comply with GDPR_ Here's a primer on anonymization and pseudonymization." Retrieved 20 JUN 2018, from https://iapp.org/news/a/looking-to-comply-with-gdpr-heres-a-primer-on-anonymization-and-pseudonymization/.
Whitehead, M., M. Suprin, T. Mistree, M. M. Kearns, G. Marini, C. Goffe, M. Pillwein and V. Abdul-Shukkoor (2024). "The Renovation of Good Clinical Practice: A Framework for Key Components of ICH E8." Therapeutic Innovation & Regulatory Science **58**(2): 303-310.
Wiktionary. "Anonymization definition." 2018, from https://en.wiktionary.org/wiki/anonymization#English.
World health Organization (WHO). "International Clinical Trials Registry Platform (ICTRP)." Retrieved 13 Mar 2024, 2024, from https://www.who.int/clinical-trials-registry-platform.
World health Organization (WHO). (2023). "Number of clinical trial registrations by location, disease, phase of development, age and sex of trial participants (1999-2022)." Retrieved 13 Mar 2024, 2024, from https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/number-of-trial-registrations-by-year-location-disease-and-phase-of-development.
World Medical Association (2013). "World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects." Jama **310**(20): 2191-2194.
Yakowitz, J. (2011). "Tragedy of the data commons." Harv. JL & Tech. **25**: 1.
Zemła-Pacud, Ż. and G. Lenarczyk (2023). "Clinical Trial Data Transparency in the EU: Is the New Clinical Trials Regulation a Game-Changer?" IIC-International Review of Intellectual Property and Competition Law **54**(5): 732-763.