Principal Investigator

First Name: Aryelly
Last Name: Rodriguez
Degree: MSc in Applied Statistics
Primary Affiliation: The University of Edinburgh
E-mail: aryelly.rodriguez@ed.ac.uk
Phone number: 07881800838
Address: Nine Edinburgh BioQuarter, 9 Little France Road
Nine Edinburgh BioQuarter, 9 Little France Road
City: Edinburgh
State or Province: Scotland
Zip or Postal Code: EH16 4UX
Country: United Kingdom

General Information

Key Personnel (in addition to PI):
First Name: Aryelly
Last name: Rodriguez
Degree: MSc in Applied Statistics
Primary Affiliation: The University of Edinburgh
SCOPUS ID: 56714457600

Are external grants or funds being used to support this research?: No external grants or funds are being used to support this research.
How did you learn about the YODA Project?: Scientific Publication

Conflict of Interest

https://yoda.yale.edu/system/files/sv_6m4tghhxg7w7uxe-r_3g6hofgxvubvnbt.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training

1. NCT00211133 - CR004414 (EPO-INT-76) DOUBLE BLIND - A Double-blind, Randomized, Placebo-controlled Study to Evaluate the Impact of Maintaining Hemoglobin Using Eprex (Epoetin Alfa) in Metastatic Breast Carcinoma Subjects Receiving Chemotherapy
2. NCT01004432 - CNTO148ART3002 - Golimumab in Rheumatoid Arthritis Participants With an Inadequate Response to Etanercept (ENBREL) or Adalimumab (HUMIRA)
3. NCT00903331 - AC-055B201 - A Double-blind, Randomized, Placebo-controlled, Multicenter, Parallel Group Study to Evaluate the Efficacy, Safety, and Tolerability of Macitentan in Patients With Idiopathic Pulmonary Fibrosis
4. NCT01715285 - 212082PCR3011 - A Randomized, Double-blind, Comparative Study of Abiraterone Acetate Plus Low-Dose Prednisone Plus Androgen Deprivation Therapy (ADT) Versus ADT Alone in Newly Diagnosed Subjects With High-Risk, Metastatic Hormone-naive Prostate Cancer (mHNPC)
5. NCT00210496 - CAPSS-334 - Efficacy of AXERT (Almotriptan Malate) in the Acute Treatment of Migraine:

A Pilot Study of the Potential Impact of Preventive Therapy With TOPAMAX (Topiramate)

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

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

Narrative Summary:

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.

Scientific Abstract:

Background; 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.
Objective; We aim to collect a broad sample of publicly available, anonymised clinical trial datasets to calculate their re-identification risk scores.
Study Design; This is a descriptive study. 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
Participants; There are not requirements on the participants to be included in this analysis, apart of being included in an anonymised/de-identified datasets from randomised controlled clinical trials.
Main Outcome Measure(s); Number of indirect identifiers present in the datasets as described by Hrynaszkiewicz et al. [3] and re-identification risk scores using all indirect identifiers in the dataset
Statistical Analysis; 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. 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.

Brief Project Background and Statement of Project Significance:

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.

For this project, we are using the following 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.

Specific Aims of the Project:

Objective
To calculate and describe the re-identification risk scores of publicly available datasets from clinical trials.

What is the purpose of the analysis being proposed? Please select all that apply.
Develop or refine statistical methods
Research on clinical trial methods

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

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
This project is not a conventional re-analysis or meta-analysis. We are interested in the re-identification risk scores for each of the selected datasets in the YODA project as they are (there would not be any pooling, imputation or manipulation of the datasets). I am not going to bring any other clinical trial datasets into the YODA project's platform from another repository/owner for analysis. Finally I understand that if I am granted access to YODA project's participant-level data, I would need to use the YODA Project remote desktop connection to a secure platform and I only will be able to export summary-level analyses. Appendix 2 in the attached study protocol has a list of all dataset repositories that we are planning to visit or visited. Unfortunately, to keep anonymity, we cannot share with YODA which datasets will be/have been extracted those repositories.

Main Outcome Measure and how it will be categorized/defined for your study:

Outcomes For each anonymised/de-identified clinical trial dataset we will calculate:
1. Number of indirect identifiers present in the datasets as described by Hrynaszkiewicz et al [3]
2. Re-identification Risk Score A (Ra)=The proportion of records that have a re-identification probability higher than 4 pre-defined thresholds (0.1 0.2 0.3 and 0.4) [13], using all indirect identifiers in the dataset

.misc-fixes { display: none; } #admin-region { z-index: 9999999; } #admin-menu { z-index: 99999999; }

3. Re-identification Risk Score B (Rb)=The Worst case scenario or weakest point in the dataset. The smallest unique group of participants (regarding all indirect identifiers in the dataset) generates the highest risk score for the whole dataset.
4. Re-identification Risk Score C (Rc) = The expected value or average risk score across all of the records in the dataset, using all indirect identifiers in the dataset
Each re-identification risk scores (A, B and C) will be estimated under the prosecutor and journalist scenario. Please see Appendix 1 in the attached protocol for more detail.

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

Number of indirect identifiers present in the datasets as described by Hrynaszkiewicz et al. [3]

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study:

Not Applicable

Statistical Analysis Plan:

Data synthesis and re-identification risk calculation
Metadata (appendix 2 in the attached study protocol) from all included datasets will be summarised in descriptive tables.
The number of indirect identifier in the anonymised/de-identified datasets will be done by visual inspection by AR and double checked by a second review (SCL or CJW). We will calculate the several re-identification risk scores with the formulas in chapter 16 from “Guide to the de-identification of personal health information” by Khaled El Eman [13] (for further details on re-identification risk scores calculations see appendix 1 in the attached study protocol) using SAS 9.4, STATA or R.

Re-identification risk scores from the anonymised/de-identified datasets will be summarised by descriptive statistics. If issues arise with any particular dataset, they will be directly discussed with the datasets owners, and if appropriate, the issue will be reported as a result in this study, anonymised and unlinked to its original anonymised/de-identified dataset. There will not be any attempt to re-identify or contact individual patients.

Data reporting and interpretation
All the re-identification risk scores in this protocol aim to assess the level of granularity in a dataset and they complement each other, so we cannot recommend one over the other. The re-identification risk scores are only driven by the number of unique indirect identifiers and the number of records in the dataset.
After calculations, we should be able to tell how datasets of a similar size with the same amount of indirect identifiers, compare to each other (e.g. datasets with 10 to 100 patients and 2-3 indirect identifiers have risk scores of around 0.3, while datasets with 100-500 patients and 2-3 indirect identifiers have risk scores of 0.2, and datasets with 500 or more patients and 2-3 indirect identifiers have risk scores of 0.1). The re-identification risk scores will be generated under both prosecutor and journalist scenarios. To help us explore their meaning, the following plots will be generated:
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.)

The re-identification risk scores by themselves cannot tell if the data has been sufficiently anonymised, but once this project is finalised they could be used to help calibrate the anonymisation process of clinical trials datasets,

.misc-fixes { display: none; } #admin-region { z-index: 9999999; } #admin-menu { z-index: 99999999; }

because they would allow dataset owners to see how much risk other researchers have taken and how theirs compares with that.

Finally, the re-identification risk scores do not have the capability of determining the probability of re-identification in the real world for a dataset. This is controlled by other factors such as: controlled vs open access to the dataset, attacker's motivations, resources and potential gains and dealing with a stigmatising intervention/disease. We will consider these factors in our final discussion when we get access to the datasets. Figure 1 in the study protocol gives an overview of the process to be followed.
Software Used:
STATA
Project Timeline:

Once access is approved and granted, we would aim to analyse and generate reports for the proposed datasets before the 31DEC2022.
We are expecting to finalise the whole project by the 30th September 2023 at the latest, at this time point all data and reports will be deleted and data contributors will be notified when the deletions has been executed.

Dissemination Plan:

Findings from this research will be presented at scientific conferences and published in a peer-reviewed journal (Clinical Trials: Sage Journal https://journals.sagepub.com/home/ctj). No publication or presentation originating from this work will reveal any data that could lead to re-identification of individuals from the data sets used.

Bibliography:

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-fu....
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-pub....
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.

Supplementary Material:

https://yoda.yale.edu/sites/default/files/rodrigueza_data_reiden_protocol_final_01_201201_.pdf
https://yoda.yale.edu/sites/default/files/cv_academic_aryelly.pdf