Evaluating re-identification risks scores in publicly available clinical trial datasets: Insights and implications

Asthma UK Centre for Applied Research

ASTHMA+ LUNG UK

aukcar.ac.uk @aukcar

Aryelly Rodriguez[1], Steff Lewis[1], Tracy Jackson[1], Christopher Weir[1], Sandra Eldridge[2]

[1]University of Edinburgh, [2]Queen Mary University of London

Introduction: 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 trial 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[1] such as date 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 adding more identifiers increases granularity by dividing the data into a larger number of smaller strata. The re-identification risk score equations[2] 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 up to one indirect identifier, it automatically scored 0 in all metrics.

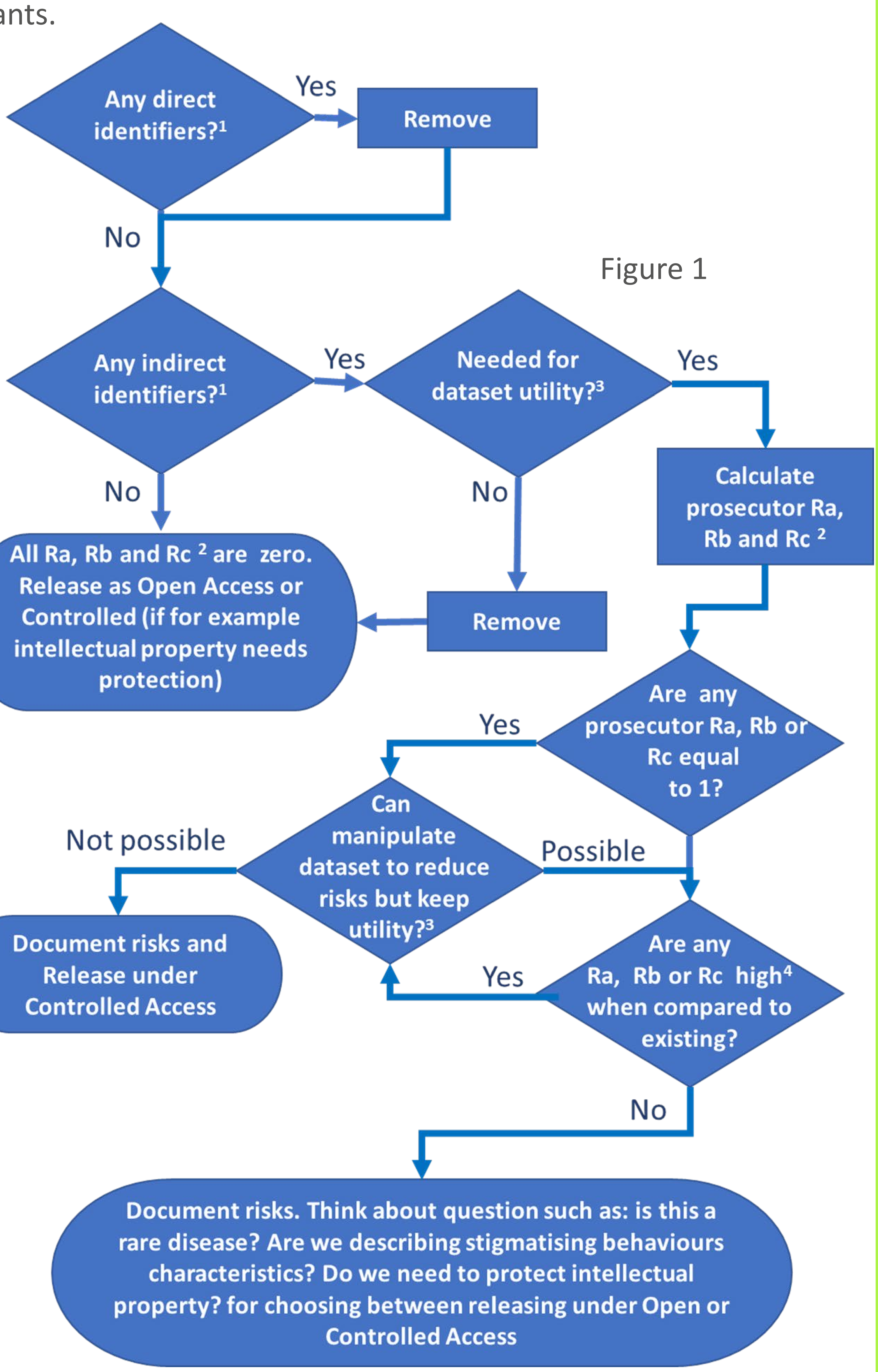

Figure 1

Any direct identifiers?[1]

Yes

Remove

No

Any indirect identifiers?[1]

Yes

Needed for dataset utility?[3]

Yes

Calculate prosecutor Ra, Rb and Rc [2]

No

All Ra, Rb and Rc [2] are zero. Release as Open Access or Controlled (if for example intellectual property needs protection)

No

Remove

Are any prosecutor Ra, Rb or Rc equal to 1?

Yes

Can manipulate dataset to reduce risks but keep utility?[3]

Not possible

Document risks and Release under Controlled Access

Possible

Are any Ra, Rb or Rc high[4] when compared to existing?

Yes

No

Document risks. Think about question such as: is this a rare disease? Are we describing stigmatising behaviours characteristics? Do we need to protect intellectual property? for choosing between releasing under Open or Controlled Access

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%).

Conclusion: 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 regarding their level of granularity prior to releasing them for secondary research. We propose a strategy for employing these scores in the decision-making process for releasing clinical trials dataset (figure 1)

[1] Hrynaszkiewicz, I., et al., Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials, 2010. 11(340).

[2] El Emam, K., Guide to the de-identification of personal health information. 2013: CRC Press.

THE UNIVERSITY of EDINBURGH

Funding: This work is funded by Asthma + Lung UK as part of the Asthma UK Centre for Applied Research (AUKCAR) [AUK-AC-2012-01 and AUK-AC-2018-01]. AR has a scholarship from the University of Edinburgh to undertake a PhD with the support from the AUKCAR. CW is funded by NHS Lothian via the Edinburgh Clinical trial Unit.

aryelly-rodriguez