Published in final edited form as:
J Public Health (Oxf). 2018 June 01; 40(2): 219–220. doi:10.1093/pubmed/fdy083.

‘Pseudonymisation at source’ undermines accuracy of record linkage

Harvey Goldstein and **Katie Harron**
UCL Great Ormond Street Institute of Child Heath, University College London,University of Bristol, School of Education, 35 Berkeley Square, Bristol BS8 1JA, United Kingdom of Great Britain and Northern Ireland

Pseudonymisation is one element of a range of measures that can be used to protect the privacy of individuals. ‘Pseudonymisation at source’ is a technique used by data providers to avoid identification of individuals before data are linked for secondary uses such as service evaluation or research. The technique involves replacement of direct identifiers, known as ‘personal data’ or ‘confidential patient information’, such as NHS number, date of birth and postcode, with a pseudonym, which does not reveal a person’s real world identity. Use of the same pseudonymisation key for multiple data sources before data are shared enables data sources to be linked together without using ‘personal data’ and therefore avoids the need for patient consent or other legal provision under the Data Protection Act or the General Data Protection Regulation.1 As we discuss below, however, this limits the utility and quality of any resulting linked datasets.

In late 2015, the Health and Social Care Information Centre, now NHS Digital (NHSD), produced its final report with recommendations on the use of ‘pseudonymisation’ in the process of linking health records such as GP records and hospital episode statistics. The group charged with producing this had spent some 3 years extensively researching the literature and formulating a set of balanced recommendations to inform NHSD policy and practice regarding pseudonymisation and considering ways to safeguard individual privacy as well as allowing the full exploitation of linked data by researchers and others. In November 2017 NHSD announced that it would not be publishing the report. Copies of the minutes of the review group can be found at2 and a response to a freedom of information request concerning reasons for non-publication can be found at.3 In March 2018, following a further freedom of information request that listed some of the key (unpublished) recommendations, NHSD did finally respond in some detail to these recommendations.4 The text of the request and the response is reproduced in the Appendix.

One of the key recommendations made by the review group concerned the use of pseudonymisation at source. The group recommended that for all new data flows into NHSD any proposal to use pseudonymisation at source would need to be fully justified, whilst

Address correspondence to Harvey Goldstein, H.Goldstein@bristol.ac.uk.

Conflict of interest
Professor Goldstein was a member of the HSCIC pseudonymisation review group.

Europe PMC Funders Author Manuscripts

existing data flows without pseudonymisation at source should continue. This recommendation has now been accepted by NHSD.4

This acceptance is important. As the review group recognized, it allows the implementation of probabilistic record linkage which is a key requirement for the development of high quality linkage across data from multiple sources in health and social care. Pseudonymisation at source means that linkage between datasets can only occur where pseudonyms agree exactly, that is, where there are no errors in the original patient identifiers used to create the pseudonym. The problem is that errors in coding identifiers occur in non-random fashion which means that failing to achieve a perfect match (agreement on pseudonyms) being associated with person characteristics.5 For example, records for patients who are socially disadvantaged may be less likely to link, which means their health needs or events, such as readmission or death, can be underestimated.5 This problem is compounded when linking more than two datasets.

NHSD is considering ways to address this problem bin the context of creating an ongoing Master Patient Services facility for linking patient records across time and environments, that will be able to use probabilistic linkage methods to improve accuracy and reduce biases in subsequent analyses using the linked data.5 If implemented this could be an important advance, augmenting the purely ‘deterministic’ procedures currently used. If data were pseudonymised at source, however, much of the benefit of such a facility would be lost by curtailing the possibilities for exploiting sophisticated linkage methods that utilize probabilistic procedures.

NHSD is the statutory body for linking health data for England. However, many data linkages take place locally using local data, often in near real-time, to plan and manage services. Increasingly, pseudonymisation at source is being used to link data from hospitals, primary care, mental health and other services for indirect health care such as commissioning or monitoring of services. Commercial companies such as MedeAnalytics are performing the pseudonymisation and the analyses, under contract to local health bodies such as Clinical Commissioning Groups (CCG).6,7 As CCGs and local authorities come to rely on these local linkages to run services and target high risk patients, it becomes increasingly important to address linkage error to avoid certain groups falling through the net. As we have suggested, pseudonymisation at source is likely to constrain the utility of any linked data. A more rational solution is to improve linkage accuracy by avoiding pseudonymisation at source, and establishing regional and national trusted linkage environments with expertise able to use more sophisticated methods, including incorporating non-disclosive patient characteristics to improve and help to evaluate linkage accuracy.8 We are encouraged by the willingness of NHSD to consider the enhancement of its own in-house expertise4 and its willingness to embrace new linkage methodologies should be an example to other groups working in this area.

Data linkers also need to publish more details about how are processed and linked,9 and the Digital Innovation Hubs envisaged in the government’s Life Sciences Strategy may be one way forward.10 Use of probabilistic methods at local, regional and national levels would

Europe PMC Funders Author Manuscripts

also enable wider linkages to data from schools and social care and could be coupled with feedback systems to improve identifier quality across sectors.

References

1. ICO anonymisation code. https://ico.org.uk/media/1061/anonymisation-code.pdf
2. Pseudonymisation review. https://www.gov.uk/government/publications/data-pseudonymisation-review#history
3. Freedom of information on non-publication. https://www.digital.nhs.uk/article/9256/Freedom-of-Information-request-NIC-156632-X6S4F
4. Reference: NIC-175129-D8K6W (not posted as of 31.03.2018)
5. Hagger Johnson: https://www.ncbi.nlm.nih.gov/pubmed/26297363
6. Fair processing of data. http://www.enhertsccg.nhs.uk/how-we-use-information-about-you-fair-processing-notice
7. Data sharing. http://www.yaxleygp.nhs.uk/sharing-your-data-61035-htm
8. https://www.ncbi.nlm.nih.gov/pubmed/29025131
9. Gilbert R, Lafferty R, Hagger-Johnson G, et al. GUILD: GUidance for Information about Linking Data sets. J Public Health. 2017; 40:191–8. DOI: 10.1093/pubmed/fdx037
10. Life sciences industrial strategy. https://www.gov.uk/government/publications/life-sciences-industrial-strategy

Europe PMC Funders Author Manuscripts