Goldstein, H., & Harron, K. (2018). 'Pseudonymisation at source' undermines accuracy of record linkage. *Journal of Public Health (United Kingdom), 40*(2), 219-220. [fdy083]. https://doi.org/10.1093/pubmed/fdy083

‘Pseudonymisation at source’ undermines accuracy of record linkage

by

Harvey Goldstein

And

Katie Harron

UCL Great Ormond Street Institute of Child Heath, University College London

Abstract

‘Pseudonymisation at source’ is the process of replacing personal identifiers with one or more pseudonymised identifiers prior to sharing or linkage of individual-level data. This technique is sometimes used to safeguard privacy and to avoid legal restrictions during the processing of data with personal identifiers. A principal disadvantage, however, is less accurate linkage and the possibility of biased results disproportionately impacting disadvantaged groups. NHS Digital has now decided not to use pseudonymisation at source on a routine basis prior to data linkage, but it is increasingly being used by local data services, and this is unnecessarily restrictive. To harness the potential of linked data for planning and managing services and research, we really need a diversity of trusted data linkage environments that can safely use identifiable, non-pseudonymised, data. Both within and outside the NHS we need suitable training in the skills to use sophisticated linkage methods, especially those associated with modern probabilistic linkage techniques to replace the restrictive ‘deterministic’ methods currently most commonly used in both the private and public sectors.

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 three 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 at (2) 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 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 recognised, 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 data sets 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 in 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 utilise probabilistic procedures.

NHS Digital 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) (See for example, 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 expertise (4) 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 also enable wider linkages to data from schools and social care and could be coupled with feedback systems to improve identifier quality across sectors.

Declaration of interest

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

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. Ruth Gilbert[1], Rosemary Lafferty, Gareth Hagger-Johnson[1], Katie Harron[2], Li-Chun Zhang[3], Peter Smith[3], Chris Dibben[4], Harvey Goldstein (2017). GUILD: GUidance for Information about Linking Data sets. Journal of Public Health | pp. 1–8 | doi:10.1093/pubmed/fdx037

10. Life sciences industrial strategy. https://www.gov.uk/government/publications/life-sciences-industrial-strategy

Appendix: NHSD response to FOI request

www.digital.nhs.uk enquiries@nhsdigital.nhs.uk

1 Trevelyan Square
Boar Lane
Leeds LS1 6AE
0300 303 5678
5th March, 2018
Our ref: NIC-175129-D8K6W

Dear Harvey Goldstein

Re: Information Request – Freedom of Information Act (FOIA) 2000

Thank you for your email dated 5th February 2018, requesting the following information:

“Thank you for this reply. I now have a further FOI request that arises from this.

You state "Instead, NHS Digital will be adopting many of the key recommendations of the Pseudonymisation Review internally alongside recommendations and requirements from the NDG Review, General Data Protection Regulation (GDPR) and other policy and legislation." I now request that you disclose precisely which recommendations NHS digital plans to adopt and when these will be implemented. Specifically I refer to the following recommendations that appeared in the final report sent to NHS digital in late 2015.

1. Pseudonymisation on its own is insufficient to protect the confidentiality of patient data The HSCIC should provide training to HSCIC staff, the wider NHS and customers which covers the organisational, legal and technical implications of using pseudonymised, data, including the risks involved and legal penalties, prior to the sharing of data
2. The HSCIC should develop an internal centre of expertise, which can provide best practice advice and guidance in relation to the ***de-identification*** *of data, including pseudonymisation for itself and the wider NHS. This would include the development of relevant standards*
3. As a priority it should:
Develop specific criteria against which individual data collections by the HSCIC can be evaluated for the optimum usage of pseudonymisation in terms of the purpose of the data collection and respecting privacy.

Develop existing techniques for anonymisation to increase the utility of the data once its disseminated

Communicate to the public the results of this activity in understandable terms.

4. Any new national data flow should be subject to IG review, through a Privacy Impact Assessment, and would involve patient groups where required. This should consider whether aggregate, fully anonymised or data pseudonymised at source could be used to meet the business objectives and realise the benefits to health and care, using data with the minimum risk of re-identification, to meet that purpose.

4. At the point that a new national data flow into the HSCIC is identified where the benefits could be fully met under a pseudonymisation at source model a Proof of Concept should be initiated to prove the efficacy of this approach in relation to the HSCIC operating model.
5. It is accepted that there are specific purposes for which the HSCIC needs to collect and process identifiable data, for example for probabilistic data linkage or clinical purposes. Existing National flows of identifiable data into the HSCIC should be subject to a rolling programme of regular review against specified criteria to ensure data flows in the least identifiable form necessary to meet the purpose. Each data flow should be reviewed in the light of legislative changes or significant technical developments, or if the requirements around individual flows change.
Please would you also confirm that, for currently processed inbound data , there is no intention intention to apply pseudonymisation at source where this currently does not exist, unless, as in 4. above a Proof of Concept is initiated to demonstrate efficacy.

Response by NHSD:

We have considered your request and in accordance with S.1 (1) of the Freedom of Information Act 2000 (FOIA) I can confirm that we do hold the information that you have requested. Please see below responses to each question.
1. Pseudonymisation on its own is insufficient to protect the confidentiality of patient data The HSCIC should provide training to HSCIC staff, the wider NHS and customers which covers the organisational, legal and technical implications of using pseudonymised, data, including the risks involved and legal penalties, prior to the sharing of data

NHS Digital has adopted much of this recommendation and plans to adopt other elements. For the planned elements, timescales are listed where planning has reached that level of detail.
NHS Digital follows the ICO Anonymisation: Code of Practice, which provides the definitive guidance on anonymisation, including the additional safeguards that must be met. Work has been undertaken to embed this across the organisation, including in the Data Access Request Service (DARS) and within the establishment of Independent Group Advising on the Release of Data (IGARD). Customers must meet stringent requirements for the sharing of data and data releases are restricted via the Data Sharing Contract and Data Sharing Agreements.
This work has been reflected within the NHS Digital published data release registers http://content.digital.nhs.uk/dataregister.
Support and guidance has also been made available to Information Asset Owners (IAOs) to help them assess the risks associated with data sharing.
Further training around data sharing will be provided to NHS Digital staff particularly as we move towards meeting General Data Protection Regulation (GDPR) in May 2018.
In the future, NHS Digital plans to publish an updated version of the Code of Practice on Confidential Information with clear requirements for the anonymisation of data, including the use of pseudonymised data. Any organisation that collects, analyses, publishes or disseminates confidential health and care information must have regard to this Code of Practice.
Relevant content will be considered for inclusion in the forthcoming Data Security and Protection Toolkit and NHS Training Tool.

2. The HSCIC should develop an internal centre of expertise, which can provide best practice advice and guidance in relation to the ***de-identification*** *of data, including pseudonymisation for itself and the wider NHS. This would include the development of relevant standards*
As a priority it should:
· Develop specific criteria against which individual data collections by the HSCIC can be evaluated for the optimum usage of pseudonymisation in terms of the purpose of the data collection and respecting privacy.
· Develop existing techniques for anonymisation to increase the utility of the data once its disseminated
· Communicate to the public the results of this activity in understandable terms.

NHS Digital plans to adopt elements of this recommendation that can be met within its budget. It plans to further develop expertise in the de-identification of data, working with other stakeholders with relevant experience and expertise. The focus of any 'centre of excellence' is planned to be on the dissemination of data rather than collection.
Over the past few years much of NHS Digital's focus has been on developing the National Data Security Centre to manage cyber security threats and incidents and to provide advice to the wider system.
In 2018, NHS Digital will look to procure a partner to develop a strategic de-identification (De-ID) solution. The key objective of this procurement is to utilise industry-leading partner expertise to enable development of both a strategic solution and to drive up internal capability in this important field.

3. *Any new national data flow should be subject to IG review, through a Privacy Impact Assessment, and would involve patient groups where required. This should consider whether aggregate, fully anonymised or data pseudonymised at source could be used to meet the business objectives and realise the benefits to health and care, using data with the minimum risk of re-identification, to meet that purpose.*

NHS Digital has already adopted much of this recommendation and plans to adopt other elements.
For new national data flows, proper review with relevant stakeholders is required in order to deliver the intended benefit in a secure and legal manner. The legal basis and information governance considerations for new flows into NHS Digital are already assured through the Data Coordination Board (DCB) process. Furthermore, where formally directed to establish a new data collection, NHS Digital would be legally obliged to collect information as specified within the Direction.
NHS Digital notes that the 'minimum risk of re-identification' for information flowing into NHS Digital does not remove the need for identifiable data to flow. As a result of NHS Digital's unique role and remit as the main safe haven for health and social care information, the majority of new national data flows are likely to include patient identifiers. In some instances, flows of anonymised or aggregate data will meet the business need e.g. Sexual and Reproductive Health Activity Data Set.

As part of changes adopted by NHS Digital to meet GDPR requirements, all existing Privacy Impact Assessments are being reviewed. If still required, these will be replaced with new Data Protection Impact Assessments (DPIAs), thus ensuring compliance with the latest data protection regulations. Where DPIAs are issued they will continue to be routinely assessed. For any new projects, systems or collections involving personal data, a DPIA will conducted. In many cases, elements will be published to support public transparency where this does not compromise security.

4. *At the point that a new national data flow into the HSCIC is identified where the benefits could be fully met under a pseuonymisation at source model a Proof of Concept should be initiated to prove the efficacy of this approach in relation to the HSCIC operating model.*

No plans have been developed on this. NHS Digital accepts that pseudonymisation at source has proved a useful mechanism, where receiving-organisations do not have a legal basis for receiving identifiable data but still need to be able to use the information.

The concept of pseudonymisation at source is already well understood and proven. If NHS Digital is directed or receives a mandatory request to establish a new system or collection using data which is pseudonymised at source, this would be subject to testing.

5. It is accepted that there are specific purposes for which the HSCIC needs to collect and process identifiable data, for example for probabilistic data linkage or clinical purposes. Existing National flows of identifiable data into the HSCIC should be subject to a rolling programme of regular review against specified criteria to ensure data flows in the least identifiable form necessary to meet the purpose. Each data flow should be reviewed in the light of legislative changes or significant technical developments, or if the requirements around individual flows change.

Elements of this recommendation have been implemented through data collections being reviewed via DPIAs.

DPIAs must be kept under review. As a minimum, the DPIA should be reviewed where changes may present a risk to the data subject's rights and freedoms. This may include changes to relevant legislature, significant technical developments or changes to the underlying requirements for a particular flow.

Please would you also confirm that, for currently processed inbound data , there is no intention to apply pseudonymisation at source where this currently does not exist, unless, as in 4. above a Proof of Concept is initiated to demonstrate efficacy.

NHS Digital confirms it has no current plans to apply pseudonymisation at source to existing collections. There are also no current plans to apply it to any new data collections, however, this is an option that would be considered when establishing any new system or data collection.

Any new systems and/or collections established by NHS Digital will be by the result of a Direction from Secretary of State or NHS England. NHS Digital will only establish new systems or collections in accordance with the stipulations of such new Directions – whether that be for the collection of data that is identifiable, pseudonymised at source or anonymised.

NHS Digital will undertake testing to demonstrate the efficacy of the new system and/or collection.

In line with the Information Commissioner’s directive on the disclosure of information under the Freedom of Information Act 2000 your request will form part of our disclosure log. Therefore, a version of our response which will protect your anonymity will be posted on the NHS Digital website.

I trust you are satisfied with our response to your request for information. However, if you are not satisfied, you may request a review from a suitably qualified member of staff not involved in the initial query, via the enquiries@nhsdigital.nhs.uk email address or by post at the above postal address.

Your request to NHS Digital will now be closed on our internal CRM (customer relationship management) system.

Yours sincerely,

J Dzakpata
Information Assurance Advisor