Making a hash of data: what risks to privacy does the NHS's care.data scheme pose?

18 March 2014

Care.Data proposes to link individual-level hospital episode statistics (HES) and general practice data at the Health and Social Care Information Centre (HSCIC). As is currently the case for HES, linked data will be pseudoanonymised before being released to researchers.[1] A proposed alternative is for identifiers (e.g. NHS number, date of birth) to be pseudoanonymised at source [2]using an encrypted hash, before linkage is performed.[3 4]

Pseudoanonymisation at source will increase data linkage errors, where two records belonging to the same patient fail to link (missed match) or two records are incorrectly assigned to the same patient (false match). Duplicate records and 'confusions' (two patients sharing a record) frequently occur in clinical settings (e.g. due to recent changes of name or address, typographical errors).

Data linkage errors have clinical implications, but are also relevant to commissioning and research. False matches lead to over-estimation of prevalence (e.g. if cases are counted twice). Missed matches lead to under-estimation of prevalence (e.g. if cases are missed) and loss of statistical power. When healthier subgroups of the population are more likely to link correctly than others, biased estimates of relative risk can occur. Linkage errors lower the quality of information available and can lead to flawed decision making.

Records that can be linked are restricted to those with complete identifiers required by the linkage algorithm. The number of these records that are correctly linked will always be lower. For example, an NHS number might be present and valid,[3] yet incorrect. Pseudoanonymisation will prevent techniques that overcome identifier errors, such as partial matching on date of birth,[1] and feedback to providers to prevent it. And if we want to plan for better integration of services across health and social care,[5] we should make best use of patient identifiers – not scramble them and ignore any errors.

Word count: 299

Corresponding author e-mail: g.hagger-johnson@ucl.ac.uk

Competing interests: GH-J has an honorary contract with the Health and Social Care Information Centre (HSCIC) as part of project funded by the Economic and Social Research Council (ESRC) to study data linkage errors. The views stated are his own.

Gareth E Hagger-Johnson, Senior Research Associate, Institute of Child Health / Department of Epidemiology and Public Health, University College London, London, United Kingdom.

Katie Harron, Research Associate, Institute of Child Health, University College London, London, United Kingdom.

Harvey Goldstein, Professor of Statistics, Institute of Child Health, University College London, London, United Kingdom.

Roger Parslow, Senior Lecturer in Epidemiology, Division of Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Leeds, United Kingdom.

Nirupa Dattani, Senior Research Fellow, Centre for Maternal and Child Health Research, City University London, London, United Kingdom.

Mario Cortina Borja, Senior Lecturer, Institute of Child Health, University College London, London, United Kingdom.

Linda Wijlaars, Research Associate, Institute of Child Health, University College London, London, United Kingdom.

Ruth Gilbert, Professor of Clinical Epidemiology, Institute of Child Health, University College London, London, United Kingdom.

References

1. HSCIC. *Replacement of the HES Patient ID (HESID)*. Leeds: Health and Social Care Information Centre, 2009.
2. Hoeksma J. The NHS's care.data scheme: what are the risks to privacy? *BMJ* 2014;348.
3. Hipisley-Cox J. Validity and completeness of the NHS number in primary and secondary care electronic data in England 1991-2013. London: University of Nottingham, 2013.
4. EMIS National User Group. EMIS NUG proposals for realising the benefits of the GP record, 2014.
5. Secretary of State. *Health and Social Care Act 2012*. London: The Stationary Office, 2010.