BMJ

BMJ 2014;348:g2264 doi: 10.1136/bmj.g2264 (Published 25 March 2014)

LETTERS

THE NHS'S CARE.DATA SCHEME

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

Gareth E Hagger-Johnson *senior research associate*[1], Katie Harron *research associate*[2], Harvey Goldstein *professor of statistics*[2], Roger Parslow *senior lecturer in epidemiology*[3], Nirupa Dattani *senior research fellow*[4], Mario Cortina Borja *senior lecturer*[2], Linda Wijlaars *research associate*[2], Ruth Gilbert *professor of clinical epidemiology*[2]

[1]Institute of Child Health/Department of Epidemiology and Public Health, University College London, London, UK; [2]Institute of Child Health, University College London, London, UK; [3]Division of Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Leeds, UK; [4]Centre for Maternal and Child Health Research, City University London, London, UK

Care.data proposes to link individual level hospital episode statistics (HES) and general practice data at the Health and Social Care Information Centre. As is currently the case for HES, linked data will be pseudoanonymised before being released to researchers.[1] A proposed alternative is for identifiers (such as 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) often occur in clinical settings (for example, owing to 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 overestimation of prevalence (if cases are counted twice). Missed matches lead to underestimation of prevalence (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, but not all of these will be correctly linked. 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 will 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.

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

Full response at: www.bmj.com/content/348/bmj.g1547/rr/689516.

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:g1547. (17 February.)
3 Hipisley-Cox J. Validity and completeness of the NHS number in primary and secondary care electronic data in England 1991-2013. University of Nottingham, 2013.
4 EMIS National User Group. EMIS NUG proposals for realising the benefits of the GP record. 2014. www.emisnug.org.uk/article/emis-nug-proposals-realising-benefits-gp-record.
5 Secretary of State. Health and Social Care Act 2012. Stationery Office, 2010.

Cite this as: *BMJ* 2014;348:g2264

g.hagger-johnson@ucl.ac.uk