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

Gareth E Hagger-Johnson senior research associate1, Katie Harron research associate2, Harvey Goldstein professor of statistics2, Roger Parslow senior lecturer in epidemiology3, Nirupa Dattani senior research fellow4, Mario Cortina Borja senior lecturer2, Linda Wijlaars research associate2, Ruth Gilbert professor of clinical epidemiology2

1Institute of Child Health/Department of Epidemiology and Public Health, University College London, London, UK; 2Institute of Child Health, University College London, London, UK; 3Division of Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, Leeds, UK; 4Centre 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. A proposed alternative is for identifiers (such as NHS number, date of birth) to be pseudoanonymised at source, using an encrypted hash, before linkage is performed.

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, yet incorrect. Pseudoanonymisation will prevent techniques that overcome identifier errors, such as partial matching on date of birth, and will feedback to providers to prevent it. And if we want to plan for better integration of services across health and social care, 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.

2 Hoeksma J. The NHS’s care.data scheme: what are the risks to privacy? BMJ 2014;348:g1547. (17 February.)

Cite this as: BMJ 2014;348:g2264

© BMJ Publishing Group Ltd 2014

g.hagger-johnson@ucl.ac.uk