Harron, K; Wade, A; Muller-Pebody, B; Goldstein, H; Gilbert, R; (2012) Opening the black box of record linkage. J Epidemiology Community Health , 66 (12) 1198. 10.1136/jech-2012-201376. Downloaded from UCL Discovery: http://discovery.ucl.ac.uk/1417397

ARTICLE

Opening the black box of record linkage

Katie Harron,[1] Angie Wade,[1] Berit Muller-Pebody,[2] Harvey Goldstein,[1] Ruth Gilbert[1]
[1]Institute of Child Health, University College London, London UK; [2]Healthcare Associated Infection, Health Protection Agency, London UK

The UK government's plan for a secure data service—*Strengthening the international competitiveness of UK life sciences research*—will transform the availability of linked electronic health records to support service provision, planning and research. In April 2012, the new Clinical Practice Research Datalink was established to provide linked national e-health records, facilitating large-scale, population-based research and service evaluation.
Such comprehensive data-linkages have been successfully established in other areas, notably in Western Australia, where a code of best practice protects patient confidentiality by separating patient-identifiable data and clinical data. Identifiers are seen only by the linkers, who provide researchers with a linked clinical dataset.

Unfortunately, linkage is not as simple as just matching identification numbers. When these are missing or inaccurate, probabilistic linkage is often used. Weights are assigned to potentially linked records, based on contributions from each partial identifier, so that agreement on name, for example, contributes more weight than agreement on sex. Only the record with the highest weight is retained and passed onto the researcher.
However, it has long been recognised that even small errors in linkage can lead to biased results.[1] Analysis from the Swiss National Cohort study showed that excluding records for which a link could not be identified underestimated cancer mortality rates.[2] Important biases also arise if linkage success differs between groups, with differential linkage by ethnic group producing biased mortality ratios.[3] Similarly, variations in data quality between hospitals can affect linkage, resulting in erroneous rankings of relative performance.[4]

The government's initiative for more effective use of existing data comes at a time when there is a lack of guidance about appropriate use of linked data, yet assessment of linkage error and its potential impacts on results can be straightforward. First, analysts can use subsets of gold-standard data where true links are known to estimate linkage sensitivity and specificity, and adjust results accordingly. Second, analysts can request measures of linkage certainty from the data linkers. For probabilistic linkage, this means providing all candidate records and associated match weights. This does not affect confidentiality, but allows comparison of the characteristics of linked and unlinked records to identify potential sources of bias, and enables sensitivity analyses using a range of linkage criteria. Finally, when current linkage methods fail to provide reasonable solutions, we need to explore alternative statistical methods—such as those using the concepts of multiple imputation—to help quantify uncertainty.[5]

We urge data-linkers to provide more from the black box of record linkage to give confidence in research arising from linked data.

References

1. Neter J, Maynes E, Ramanathan R. The effect of mismatching on the measurement of response error. J Am Stat Assoc 1965;60:1005e27.
2. Clough-Gorr K. The Challenge of Unlinked Deaths in Health Research: an Example from the Swiss National Cohort Study. Exploiting Existing Data for Health Research, Scottish Health Informatics Programme St Andrews. 2011.
3. Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality Studies. J Aging Health 2011;23:1263e84.
4. Gibbs JL, Cunningham D, De Leval M, et al. Paediatric cardiac surgical mortality after Bristol. BMJ 2005;330:43e4.
5. McGlincy M, ed. A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links. ASA Section on Survey Research Methods. 2004. http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000683.pdf (accessed 6 Mar 2012).