ARTICLE

Opening the black box of record linkage

Katie Harron,1 Angie Wade,1 Berit Muller-Pebody,2 Harvey Goldstein,1 Ruth Gilbert1
1Institute of Child Health, University College London, London UK; 2Healthcare
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