**E**ach of us, in living our lives, creates a vast amount of data. Medical records, from the moment we are born. Education records, from the time our schooling begins. Tax and income records, beginning with our first payslip.

Such records are valuable to researchers who seek to understand the pathways people take through life. But there is a catch: data about the same persons are often stored in different places, in different organisations. So, to get the most out of this information, these separate records need to be linked.

The term "record linkage" was first and perhaps best described in 1946 by Halbert Dunn, chief of the National Office of Vital Statistics in the US Public Health Service.[1] He said: "Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume."

Record linkage is conceptually simple – find all records with matching names, dates of birth or other identifiers. However, it is much more complex in practice, especially when dealing with population-level data, in which tens of millions of records might need to be linked for hundreds of thousands of people. Clearly, such record linkage needs to be automated – but automated systems are not foolproof. For example, data from the same person might not be matched if names have been recorded differently, perhaps due to user error.

The methods underpinning much of the linkage performed today are based on simple statistical concepts proposed in the 1950s by the geneticist Howard Newcombe,[2] and formalised in 1969 by Ivan Fellegi and Alan Sunter of the Dominion Bureau of Statistics (a Canadian government organisation responsible for censuses).[3] Their idea was that linkage could be automated by creating weights that represent how likely it is that two records belong to the same entity. These "match weights" are used to classify pairs of records into M (a set of true matches) and U (a set of true non-matches),

based on the ratio of two probabilities:

$R = P(\gamma|M)/P(\gamma|U)$

where $\gamma$ represents the agreement pattern between two records. The agreement pattern denotes the extent to which a common set of identifiers agree for a particular pair of records. This might be based on simple agreement or disagreement, distance measures (e.g., string similarity comparators), or it may take into account the relative frequency with which specific values of identifier values occur in the data (e.g., allowing that "Harron" would generally be a less common surname than "Smith").
The idea is that match weights take into account both accuracy and discriminatory power of identifiers: the parameter $P(\gamma|M)$ is known as the "m-probability" and is related to the accuracy of recorded identifiers (e.g., name might be more prone to typographical errors than date of birth); $P(\gamma|U)$ is known as the "u-probability" and relates to the discriminatory power of a particular identifier (e.g., postcode is more discriminatory than sex). The final match weight is derived as a function (usually the binary logarithm) of the ratio of probabilities, summed across identifiers. The result is that records belonging to the same individual should be represented with high weights, and those belonging to different individuals should be represented by low weights. Agreement on highly discriminative identifiers increases the size of the weight more than agreement on low discriminative identifiers; disagreement on poor quality identifiers penalises the weight less than disagreement on high quality identifiers.
This type of linkage approach is often described as "probabilistic" – although, in practice, a single threshold weight is usually used to classify pairs of records as matches or not. In this sense, probabilistic linkage is a highly flexible extension to simpler deterministic (or rule-based) linkage approaches that classify record pairs based on a defined set of agreement patterns.

Data linkage has become an increasingly important tool for research, and for building strong evidence on which to base decisions on public policy. Linkage across generations or households, for example, is providing new insights into how environmental, genetic and social factors in early life might influence later health and development. Linkage can also be used to transform the scope, design and efficiency of primary studies such as clinical trials, surveys and cohorts, and has been key to providing rapid evidence on the impact of the Covid-19 pandemic.

Still, a number of challenges remain. Firstly, perfectly accurate and unique identifiers are unlikely ever to exist in the types of data sources we would like to link (at least while human error has a part to play). However, getting linkage right is crucial if we want to produce reliable results with which to inform public policy. Even small levels of error in linkage can lead to substantially biased results, particularly if those errors predominate in specific subgroups (such as ethnic groups) that might be of interest for analysis. Optimising linkage is a balance between investing in improved data capture and implementation of linkage methods, and clear communication of uncertainty in linkage so that analysts can employ statistical methods to account for error.

In order for us to fully realise the potential of linked data, balancing privacy (for the individual) and quality (for research) is key. Promoting public trust in the use of linked data is also crucial for us to continue using these data to inform evidence based decisions on public policy.

Disclosure statement

The author holds research grant funding from the Wellcome Trust and the National Institute for Health Research.

References

1. Dunn, H. L. (1946) Record linkage. American Journal of Public Health, **36**(12), 1412–1416.

2. Newcombe, H. B., Kennedy, J. M., Axford, S. J. and James, A. P. (1959) Automatic linkage of vital records.

Science, **130**(3381), 954–959.
3. Fellegi, I. P. and Sunter, A. B. (1969) A theory for record linkage. Journal of the American Statistical Association, **64**(328), 1183–1210.