bmjmedicine

Check for updates

Data linkage in medical research

Katie Harron

Data linkage provides an opportunity to harness existing data for medical research. This article outlines key approaches for data linkage, and describes methods used to quantify, interpret, and account for errors.

Correspondence to: Dr Katie Harron, UCL Great Ormond Street Institute of Child Health, Population Policy and Practice, London, UK; k.harron@ucl.ac.uk

Cite this as: *BMJMED* 2022;1:e000087. doi:10.1136/bmjmed-2021-000087

Received: 9 December 2021
Accepted: 13 January 2022

Data linkage combines data from different sources that relate to the same person to create a new, enhanced data resource. This technique allows researchers to exploit and enhance existing data sources without the time and cost associated with primary data collection. Linked data can be used to supplement follow-up in conventional cohort studies or trials, or to generate real world evidence by creating population level electronic cohorts that are entirely derived from administrative data (figure 1).[1 2] These longitudinal data sources help us to answer questions that require large sample sizes (eg, for rare diseases) or whole population coverage (eg, for pandemic response planning), which consider a wide range of risk factors and outcomes (including social determinants) and are especially powerful for capturing populations that are hard to reach.[3 4] Figure 1 illustrates two real world examples of how data linkage has been used to inform medical research, to improve clinical trial follow-up and to examine outcomes from birth.

Choosing the right approach

A barrier to generating linked data that are fit for purpose is the availability of accurate identifiers that can be used to link the same person across multiple data sources.[5] Recording of unique identifiers such as the NHS number nearly always involves some degree of error or missing data.[6] Therefore, linkage often depends on the use of non-unique identifiers such as name, postcode, and date of birth, or even indirect identifiers such as procedure dates or other clinical variables.[7] In combination, these variables can allow us to identify records that belong to the same person—but errors, changes over time, or missing data can still hamper attempts to find the correct link.

Linkage methods have two main categories: deterministic (rule based) methods and probabilistic methods involving match weights or scores. For example, national hospital records in England (Hospital Episode Statistics) are linked longitudinally for the same person using a three step deterministic algorithm that looks for exact agreement on a combination of identifiers: NHS number, date of birth, postcode, and sex.[8] Probabilistic linkage assigns a weight to each record pair, representing the likelihood that two records belong to the same individual, given the agreement or not between identifiers. In effect, probabilistic match weights allow all possible deterministic rules for a set of available identifiers to be ranked.[9] A threshold weight is then used to classify records as links or non-links.

Linkage error

Irrespective of the linkage methods implemented, use of imperfect and dynamic identifiers can lead to linkage error. Linkage errors manifest as false matches (where records belonging to different individuals are linked together) or missed matches (where records belonging to the same individual are not linked). Analogous to false positives and false negatives, these linkage errors can be viewed through a diagnostic accuracy lens (table 1). While carefully designed linkage algorithms and high quality recording of identifying information can facilitate accurate linkage, even small amounts of error can lead to bias.[10] This problem is particularly evident when individuals from certain subgroups are less likely to link accurately.[11] For example, maintaining consistent linkage quality across ethnic groups can be a challenge.[12]

Any linkage strategy will allow, to a certain extent, a trade-off between the two types of errors.[9] In probabilistic linkage, this trade-off depends on the choice of threshold (that is, the weight above which records have been classified as links; figure 2). As the threshold is lowered, sensitivity of linkage (that is, the proportion of true matches captured) increases, but the false match rate also increases. A similar diagram could be drawn to represent the trade-off in deterministic linkage as we decide which matching rule or match rank should be used to classify records. Sensitivity analyses can be used to explore the impact of the choice of threshold or matching rule on results.[13]

Design of a linkage strategy should be informed by the intended application or research question. For example, when creating a system to support drug administration using linked records, we would need to ensure that treatments are not delivered to the wrong patient: a conservative or specific approach aiming to minimise false matches would be appropriate. Conversely, use of linked data to invite members of the public for screening programmes might prioritise coverage at the expense of sending some invitations in error: a more sensitive approach might be appropriate in this setting. Minimising the

KEY MESSAGES

⇒ Data linkage in medical research allows researchers to exploit and enhance existing data sources without the time and cost associated with primary data collection

⇒ Methods used to quantify, interpret, and account for errors in the linkage process are needed, alongside guidelines for transparent reporting

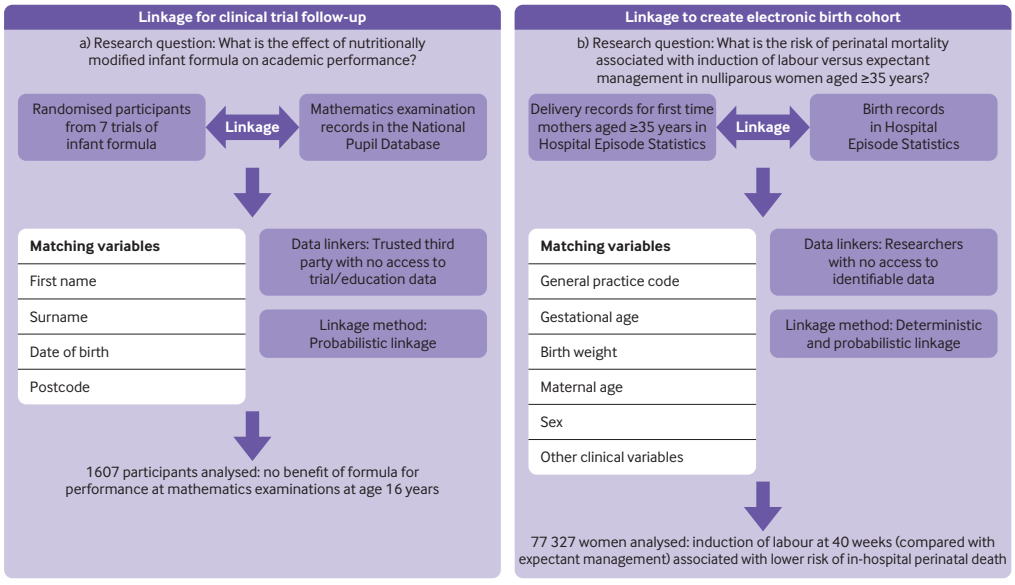

Figure 1 | Examples of linkage used to support clinical trials and create whole population cohorts.[21 22]

difference between error types might also important in some situations. For example, when mortality rates are estimated by linking a cohort to mortality records, the correct rate might still be estimated if the number of false and missed matches cancel out.

Quality control and accounting for linkage error

Several methods can be used to evaluate the quality of linkage.[14] These methods focus on identifying potential sources of bias (that is, which characteristics are associated with errors) by examining the characteristics of records that are linked versus unlinked, or that have high versus low quality identifier data, or that are easily identifiable as having been linked incorrectly (eg, through quality control checks).[15] Accounting for linkage error in analysis is an ongoing area of methodological research, but includes approaches that view uncertainty in linkage as a missing data problem best handled with some form of multiple imputation or weighting, and those that attempt to quantify and adjust for errors using quantitative bias analysis.[16] Reporting guidelines are available that explicitly aim to support transparent reporting of linkage studies.[5 17]

Table 1 | Linkage accuracy tool

	True match status	
Assigned link status	**Match (pair from same individual)**	**Non-match (pair from different individuals)**
Link	True match a	False match b
Non-link	Missed match c	True non-match d

Sensitivity (or recall)=a/(a+c); specificity=d/(b+d); positive predictive value (or precision)=a/(a+b); negative predictive value=d/(c+d). In practice, the number of non-matches will usually far outweigh the number of matches, and so the positive predictive value and sensitivity are more informative than the specificity and negative predictive value.

Remaining challenges

The biggest barriers to realising the full potential of data linkage as a powerful research tool are gaining and maintaining public trust, and reducing the costs, delays, and inefficiencies in how linked data are made available for research in the public interest.[18 19] For example, proposals to routinely link health records in primary and secondary care in order to support planning and research in England (from care.data in 2012 to General Practice Data for Planning and Research in 2021) have repeatedly raised public concerns about the lack of transparency surrounding how linked data are to be used, processes for opting out, and commercial interests. However, the covid-19 pandemic has highlighted that efficient and secure access to linked data can support agile and responsive research: building on the success of initiatives such as OpenSafely and the

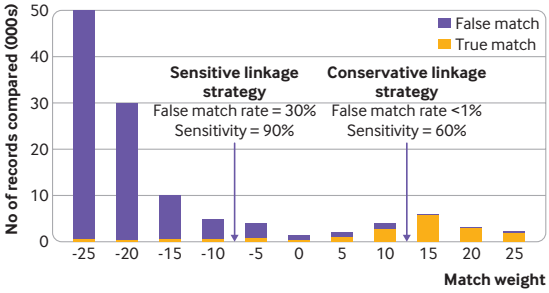

Figure 2 | Example of trade-off between false matches and missed matches in probabilistic linkage. In this example, probabilistic match weights are used to classify records as belonging to the same individual or not. A threshold of ≥15 would mean that <1% of linked records were false matches but 40% of the true matches were not captured. Decreasing the threshold to −5 would increase the proportion of true matches captured to 90%, but would also increase the false match rate to 30%

Harron K. *BMJMED* 2022;1:e000087. doi:10.1136/bmjmed-2021-000087

British Heart Foundation's CVD-COVID-UK consortium (both of which link primary and secondary care data for the UK) could provide a way forward.[1 20]

Contributors KH wrote the article.

Competing interests We have read and understood the BMJ policy on declaration of interests and declare the following interests: none.

Patient and public involvement Patients and the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Provenance and peer review Commissioned; not externally peer reviewed.

Data availability statement No data are available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

ORCID iD

Katie Harron http://orcid.org/0000-0002-3418-2856

REFERENCES

1 Wood A, Denholm R, Hollings S, *et al*. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021;373:n826. doi:10.1136/bmj.n826
2 Fitzpatrick T, Perrier L, Shakik S, *et al*. Assessment of long-term follow-up of randomized trial participants by linkage to routinely collected data: a scoping review and analysis. *JAMA Netw Open* 2018;1:e186019. doi:10.1001/jamanetworkopen.2018.6019
3 Chiu M, Lebenbaum M, Lam K, *et al*. Describing the linkages of the immigration, refugees and citizenship Canada permanent resident data and vital statistics death registry to Ontario's administrative health database. *BMC Med Inform Decis Mak* 2016;16:135. doi:10.1186/s12911-016-0375-3
4 Aldridge RW, Menezes D, Lewer D, *et al*. Causes of death among homeless people: a population-based cross-sectional study of linked hospitalisation and mortality data in England. *Wellcome Open Res* 2019;4:49. doi:10.12688/wellcomeopenres.15151.1
5 Gilbert R*et al*. Guild: guidance for information about linking datasets. *J Public Health* 2017:1–8.
6 Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, *et al*. The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *Eur J Epidemiol* 2009;24:659–67. doi:10.1007/s10654-009-9350-y
7 Blake HA, Sharples LD, Harron K, *et al*. Probabilistic linkage without personal information successfully linked national clinical datasets. *J Clin Epidemiol* 2021;136:136–45. doi:10.1016/j.jclinepi.2021.04.015
8 Hagger-Johnson G, Harron K, Goldstein H, *et al*. Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. *J Innov Health Inform* 2017;24:234. doi:10.14236/jhi.v24i2.891
9 Doidge JC, Harron K. Demystifying probabilistic linkage. *Int J Popul Data Sci* 2018;3. doi:10.23889/ijpds.v3i1.410
10 Harron K, Dibben C, Boyd J, *et al*. Challenges in administrative data linkage for research. *Big Data Soc* 2017;4:205395171774567. doi:10.1177/2053951717745678
11 Bohensky MA, Jolley D, Sundararajan V, *et al*. Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010;10:346–52. doi:10.1186/1472-6963-10-346
12 Grath-Lone LM, Libuy N, Etoori D, *et al*. Ethnic bias in data linkage. *Lancet Digit Health* 2021;3): :e339. doi:10.1016/S2589-7500(21)00081-9
13 Harron KL, Doidge JC, Knight HE, *et al*. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol* 2017;46): :1699–710. doi:10.1093/ije/dyx177
14 Doidge J*et al*. *Quality assessment in data linkage. Joined up data in government: the future of data linking methods*. London: Office for National Statistics, 2020. https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage
15 Doidge J, Harron K. Linkage error bias. *Int J Epidemiol* 2019:dyz203.
16 Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. *Ann Hum Biol* 2020;47): :218–26. doi:10.1080/03014460.2020.1742379
17 Benchimol EI, Smeeth L, Guttmann A, *et al*. The reporting of studies conducted using observational Routinely-collected health data (record) statement. *PLoS Med* 2015;12:e1001885. doi:10.1371/journal.pmed.1001885
18 Taylor JA, Crowe S, Espuny Pujol F, *et al*. The road to hell is paved with good intentions: the experience of applying for national data for linkage and suggestions for improvement. *BMJ Open* 2021;11:e047575. doi:10.1136/bmjopen-2020-047575
19 Cavallaro F*et al*. Reducing barriers to data access for research in the public interest - lessons from covid-19. *BMJ Opinion* 2020.
20 Mathur R, Rentsch CT, Morton CE, *et al*. Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform. *Lancet* 2021;397:1711–24. doi:10.1016/S0140-6736(21)00634-6
21 Verfürden ML, Gilbert R, Lucas A, *et al*. Effect of nutritionally modified infant formula on academic performance: linkage of seven dormant randomised controlled trials to national education data. *BMJ* 2021;375:e065805. doi:10.1136/bmj-2021-065805
22 Knight HE, Cromwell DA, Gurol-Urganci I, *et al*. Perinatal mortality associated with induction of labour versus expectant management in nulliparous women aged 35 years or over: an English national cohort study. *PLoS Med* 2017;14:e1002425. doi:10.1371/journal.pmed.1002425