A celebration of Harvey Goldstein's lifetime contributions

A journey in data linkage with Harvey Goldstein

Katie Harron
UCL Great Ormond Street Institute of Child Health

Where it all started

I began working with Harvey as a PhD student at ICH in 2010. My project was part of a large clinical trial (CATCH) aiming to evaluate whether an intervention in paediatric intensive care units (PICUs) – i.e., central lines that were impregnated with antibiotics – was effective at reducing the risk of bloodstream infection. My role was to analyse background trends in infection rates in PICUs to understand the generalisability of the results of the trial to units across the country. The challenge was that no single data source reliably captured information nationally on both admissions to PICU and bloodstream infection. To accurately estimate rates of infection for children on PICU, we needed to link two data sources: PICANet (the Paediatric Intensive Care Audit Network dataset, which captures information about all admissions to PICU in England) and LabBase2 (infection surveillance data collated by Public Health England at the time).[1]

As someone new to the field I expected this linkage to be a straightforward task, relying on the National Health Service number being recorded accurately and completely in both data sources. However, we soon realised that it was a bit more complicated than that. NHS number was only complete for around 50% of records in the infection surveillance system, and so we needed to consider other non-unique identifiers such as name, date of birth, sex, postcode, and hospital location. The completeness of these identifiers ranged from 22% (first name) to 96% (sex).

The problem with the use of these non-unique identifiers is the potential for introducing linkage errors. In the context of measuring infection rates, missed links (where an infection record should have been linked to a PICANet record, but wasn't) can lead to an underestimated infection rate. False links (where an infection record was erroneously linked to a PICANet record) can overestimate the infection rate. A further complication was trends over time: improvements in data quality over time meant that records were more likely to link correctly in

more recent years than earlier years, which could lead to the false impression that infections rates were increasing (or staying stable, if in fact they were decreasing).

Re-framing the problem

Harvey's key insight to this problem was that rather than focussing on creating one 'perfect' linkage solution, we should be concerned with correctly representing the uncertainty in the analysis. He suggested that rather than linking *records*, we should think about linking data values. In other words, the emphasis shifts to obtaining correct estimates for model parameters in the analysis of interest. In our case, he meant that we should aim to carry over the correct value of our variable of interest (infection or no infection) to the analysis. In this way, Harvey reframed the linkage problem using a missing data framework.

This idea was based on a paper Harvey had written with colleagues in 2009 on multilevel models with multivariate mixed response types.[2] In this paper, he described "partially observed" data, where there is some uncertainty about the correct value of a particular variable, but some information about the possible candidate values. He suggested a method for handling partially observed data, called "prior-informed imputation", which extends the standard multiple imputation framework by allowing the inclusion of an informative prior distribution.

Figure 1 gives an example of a set of records to be linked: some will be linked with certainty, whilst others have no candidate links and will remain unlinked. The remainder have equivocal links, i.e., they are associated with one or more candidate linking records, but there is some uncertainty about which, if any, is the correct link. We have assigned a measure of linkage certainty, in this case, using probabilistic match weights. Probabilistic match weights are typically calculated by combining information on the accuracy with which identifiers are recorded (the m-probability, i.e., the probability that a particular identifier agrees in records belonging to the same people), and the discriminatory power of an identifier (the u-probability, i.e., the probability that a particular identifier agrees in records belonging to different people).[3]

If we only included those records in Figure 1 that are linked or not linked with certainty, this is analogous to a complete case analysis, and would result in a smaller sample size, a loss of statistical power, and potential bias (especially if linkage certainty is related to the outcome).[4] Alternatively, we could consider the uncertain links to be missing data and impute the variable of interest using standard multiple imputation procedures, taking into account any auxiliary variables.[5] However, this approach would not make use of all the information we have – namely, we know something about the likelihood of each potential value of the variable of interest from our probabilistic match weights. This is how Harvey applied the idea of prior-informed imputation to linkage: the probabilistic match weights form the prior distribution.[6]

Taking a step back from the conventional ways of doing things, and reframing the problem to fit with existing approaches, was a real strength of Harvey's insight. Incidentally, he also challenged the predominant approach to calculating match weights that had originally been formalised in the 1960s by Fellegi and Sunter, highlighting that the approach was not grounded in statistical theory.[7, 8] He developed an alternative approach by returning to his earlier work on correspondence analysis of wrist bone maturity development in children.[9]

Thinking outside the black box

As Harvey became more interested in the intricacies of data linkage and linkage error, and the potential for linkage of administrative data from government departments for generating evidence with high external validity, he became a strong advocate of transparency.[10] Due to the sensitive nature of the data required for linkage, much of the linkage of administrative data in the UK is performed in what feels like a black box, within data holder organisations or "trusted third parties", rather than by researchers.[11] Through persistent engagement with the Health and Social Care Information Centre (now NHS Digital), Harvey contributed to the design of their linkage strategies through their Data Linkage Advisory group, organised for a UCL researcher to have a placement within their linkage department in order to evaluate existing linkage methods, and was vocal in his misgivings about plans for removing identifiers prior to linkage of health data (pseudonymisation at source).[12-15] He was also imaginative and innovative in his thinking about how to handle tricky linkage problems. For example, his ideas about using non-traditional identifiers for linkage, such as height and weight recorded in different datasets, led to the development of linkage between mother and baby hospital records for the whole population of England.[16] These linked data have been used to support a number of studies investigating the impact of maternal exposures on later child outcomes.[17-20]

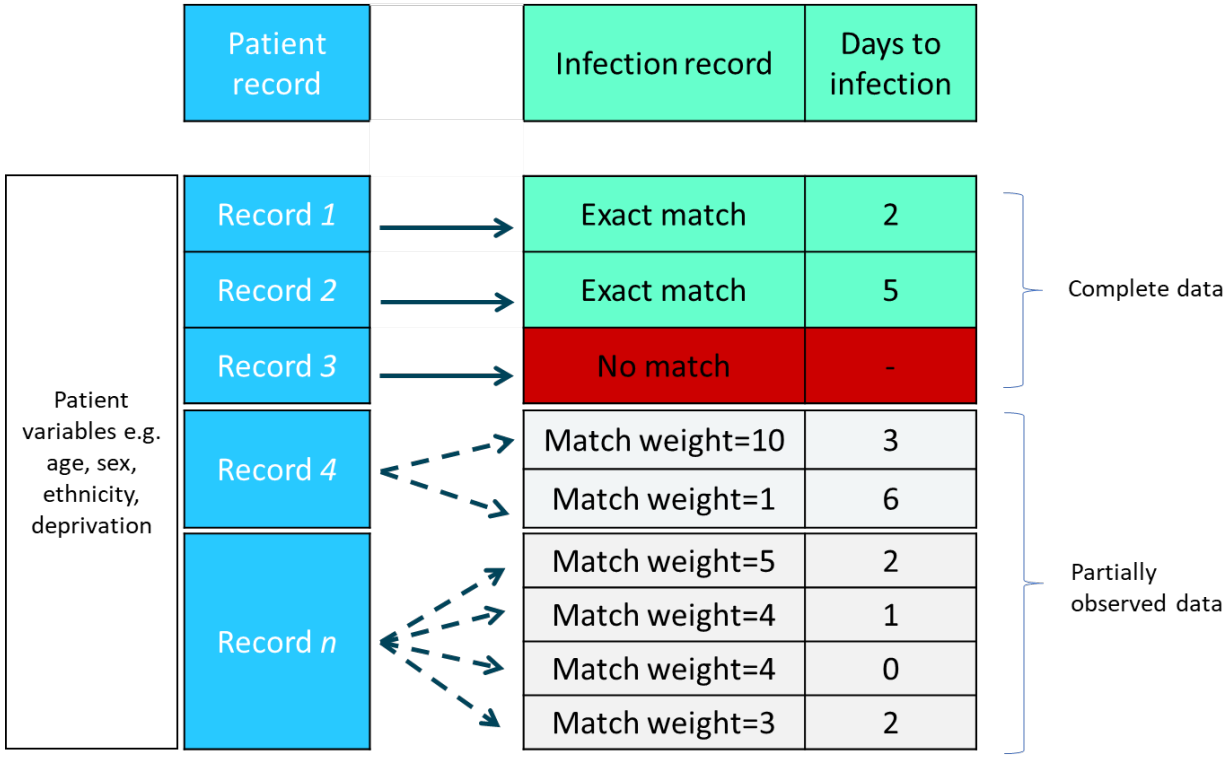
Harvey was very effective in engaging with the right stakeholders in order for his ideas to have impact, and jointly wrote guidance on what information should be shared about linkage, with colleagues at UCL and the Office for National Statistics.[21] Although just one of his varied interests at the time, Harvey became a leading authority on data linkage in the UK, he was commissioned to edit a Wiley textbook on methodological developments in the field, and was contributing to the National Statistician's cross-government review of data linkage methods right up until his death in 2020.[22, 23]

A linkage legacy

I will remember Harvey primarily for his generosity in the time that he gave to support me and other junior colleagues, for his approachability despite having such a formidable reputation, and for his sense of humour and joy in his work. I hope to learn from his perseverance in finding solutions and bringing them to the right people, and for constantly challenging the

status quo. As we rely more and more on linked data for research, service planning and public policy, we will continue to benefit from Harvey's contribution to data linkage methods.

Figure 1: Partially observed data in linkage between a patient file and an infection file. "Days to infection" is the variable of interest, to be analysed with patient variables held in the patient file. Patient records 1 and 2 have a certain link in the infection file; record 3 has no link (this patient had no infection). Record 4 and record n have several possible links, and are therefore partially observed. Values for these records could be imputed using standard multiple imputation, or using prior-informed imputation in which the match weights are used to form a prior distribution for the variable of interest.



References

1. Harron, K., et al., *Linkage, evaluation and analysis of national electronic healthcare data: application to providing enhanced blood-stream infection surveillance in paediatric intensive care.* PLoS One, 2013. **8**(12): p. e85278.
2. Goldstein, H., et al., *Multilevel models with multivariate mixed response types.* Stat Model, 2009. **9**(3): p. 173-197.
3. Doidge, J. and K. Harron, *Demystifying probabilistic linkage.* Int J Popul Data Sci, 2018. **3**(1).
4. Harron, K., J.C. Doidge, and H. Goldstein, *Assessing data linkage quality in cohort studies.* Annals of Human Biology, 2020. **47**(2): p. 218-226.
5. Harron, K., et al., *Evaluating bias due to data linkage error in electronic healthcare records.* BMC Med Res Methodol, 2014. **14**(1): p. 36.
6. Goldstein, H., K. Harron, and A. Wade, *The analysis of record-linked data using multiple imputation with data value priors.* Stat Med, 2012. **31**(28): p. 3481-93.
7. Goldstein, H., K. Harron, and M. Cortina-Borja, *A scaling approach to record linkage.* Stat Med, 2017. **36**: p. 2514-21.
8. Fellegi, I. and A. Sunter, *A theory for record linkage.* J Am Stat Assoc, 1969. **64**(328): p. 1183-1210.
9. Goldstein, H., *The choice of constraints in correspondence analysis.* Psychometrika, 1987. **52**(2): p. 207-215.
10. Goldstein, H., *Living by the evidence.* Significance, 2020. **17**(1): p. 38-40.
11. Harron K, et al., *Opening the black box of record linkage.* J Epidemiol Commun H, 2012. **66**(12): p. 1198.
12. Hagger-Johnson, G., et al., *Identifying false matches in anonymised hospital administrative data without patient identifiers* Health Serv Res, 2014. **50**(4): p. 1162-78.
13. Hagger-Johnson, G.E., et al., *Making a hash of data: what risks to privacy does the NHS's care.data scheme pose?* BMJ, 2014. **348**(g2264).

14. Hagger-Johnson, G., et al., *Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data.* J Innov Health Inform, 2017. **24**(2): p. 891.
15. Goldstein, H. and K. Harron, *'Pseudonymisation at source' undermines accuracy of record linkage.* Journal of Public Health, 2018: p. fdy083-fdy083.
16. Harron, K., et al., *Linking data for mothers and babies in de-identified electronic health data.* PLoS One, 2016. **11**(10): p. e0164667.
17. Knight, H.E., 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**(11): p. e1002425.
18. Harron K, et al., *Associations between pre-pregnancy psychosocial risk factors and infants outcomes: a population-based cohort study in England.* Lancet Public Health, 2021. **6**(2): p. e97-105.
19. Guttmann, A., et al., *Long-term mortality in mothers of infants with neonatal abstinence syndrome: A population-based parallel-cohort study in England and Ontario, Canada.* PLOS Medicine, 2019. **16**(11): p. e1002974.
20. Zylbersztejn, A., et al., *Child mortality in England compared with Sweden: a birth cohort study.* Lancet, 2018. **391**(10134): p. 2008-2018.
21. Gilbert, R., et al., *GUILD: Guidance for Information about Linking Datasets.* J Public Health, 2017. **1-8**.
22. Harron, K., C. Dibben, and H. Goldstein, *Methodological developments in data linkage*. 2015: Wiley.
23. Goldstein H, *Efficient procedures for linking datasets for the purpose of fitting statistical models*, in *Joined up data in government: the future of data linking methods* Office for National Statistics, Editor. 2020: London.