International Journal of Population Data Science

Journal Website: www.ijpds.org

Demystifying probabilistic linkage: Common myths and misconceptions

Doidge, JC[1,2*] and Harron, K[3]

Submission History

Submitted:	03/08/2017
Accepted:	07/12/2017
Published:	10/01/2018

[1]Administrative Data Research Centre for England, University College London
[2]Centre for Population Health Research, University of South Australia
[3]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine

Abstract

Many of the distinctions made between probabilistic and deterministic linkage are misleading. While these two approaches to record linkage operate in different ways and can produce different outputs, the distinctions between them are more a result of *how* they are implemented than because of any intrinsic differences. In the way they are generally applied, probabilistic and deterministic procedures can be little more than alternative means to similar ends—or they can arrive at very different ends depending on choices that are made during implementation. Misconceptions about probabilistic linkage contribute to reluctance for implementing it and mistrust of its outputs. We aim to explain how the outputs of either approach can be tailored to suit the intended application, but also to highlight the ways in which probabilistic linkage is generally more flexible, more powerful and more informed by the data. This is accomplished by examining common misconceptions about probabilistic linkage and its difference from deterministic linkage, highlighting the potential impact of design choices on the outputs of either approach. We hope that better understanding of linkage designs will help to allay concerns about probabilistic linkage, and help data linkers to select and tailor procedures to produce outputs that are appropriate for their intended use.

Keywords

medical record linkage; data linkage; data matching; electronic health records; probabilistic linkage; deterministic linkage; record linkage

Introduction

Many of the distinctions made between probabilistic and deterministic linkage are misleading. This seems to be largely because of misinterpretation of the 'match weights' that lie at the heart of the method first formalised by Fellegi and Sunter (1). There are some important differences between the procedures, but these bear little relation to the distinctions frequently drawn. Adoption of probabilistic linkage has been haphazard, with some data linkers—often from research backgrounds—promoting probabilistic techniques, while others—often government service providers-have been more hesitant. Service provision and research are very different applications of linked data and can require different outputs. In this article we aim to explain how the outputs of either approach can be tailored to suit the intended application, but also to highlight the ways in which probabilistic linkage is generally more flexible, more powerful and more informed by the data.

While there are settings in which a simpler, deterministic approach to linkage can be sufficient, and there are settings in which the standard assumptions used to generate match weights may not be appropriate (2), there are generally no settings in which the quality of a deterministic linkage cannot be equalled or exceeded by a well-designed probabilistic linkage. Reluctance against implementing probabilistic linkage and poor understanding of how each procedure can be tailored to the application therefore undermine the quality of data linkage and the services and research that rely on it.

Some analysts have attempted to draw empirical comparisons of probabilistic and deterministic linkage (3-5) but this is arguably misguided, as the performance of both techniques depends critically on a number of decisions that are made about how each is implemented (6). This paper aims to improve readers' understanding of how record linkage procedures operate and how they can be tuned and adapted to produce quite similar—or very different—outputs, depending on the objectives of the data linker. To achieve this aim, we present a critical discussion of some common myths and misconceptions about probabilistic linkage and the differences between probabilistic and deterministic linkage. We shall begin with a brief introduction to record linkage techniques. A more detailed introduction to probabilistic record linkage is provided by Sayers, Ben-Shlomo (7) and a thorough overview by Winkler (8). An overview of recent developments in data linkage in general is provided in Harron, Goldstein (9).

*Corresponding Author:
Email Address: j.doidge@ucl.ac.uk (JC Doidge)

http://dx.doi.org/10.23889/ijpds.v3i1.410

Background: Deterministic and probabilistic approaches to record linkage

There are many applications for record linkage, with terminologies that have evolved simultaneously in several fields. Data linkage, data matching, record linkage, record matching, merging, entity resolution, deduplication and reidentification can all mean the same thing. The essence of the problem is comparing pairs of records (observations or rows in one or multiple files) to identify whether they relate to the same or different unit of observation (entity, person, etc.). Records are compared using a set of one or more matching variables (identifiers) that are common to both records.

'Deterministic linkage' involves setting decision rules based on agreement on matching variables (e.g. 'records agree on social security number' or '. . . on first name, surname and date of birth'). Joining tables using a single unique 'key' variable (e.g. NHS or social security number) is the simplest example of deterministic linkage. The problem becomes non trivial, however, when there is no unique linkage key and available matching variables lack uniqueness, contain errors, or are missing. In these settings, multiple decision rules will often be specified and implemented sequentially, usually starting with those that are thought to be least likely to return 'false matches' (links between records belonging to different entities). Less specific rules are then applied, with the aim of detecting more matches but with increasing risk of false matches. When a series of decision rules are applied, the step or 'match rank' at which a link is identified can be used as an indicator of confidence or uncertainty about the link. Selecting and ranking decision rules can be difficult and is often based on the subjective intuition of the linker. This is the main problem that probabilistic linkage aims to address.

'Probabilistic linkage' uses statistical theory to associate each pattern of matching variable agreement with the likelihood that record pairs exhibiting the pattern are a match. It does this using two sets of probabilities: the probability that records agree on each matching variable given that the records truly are a match (the m-values) and the probability that they agree if they are truly a nonmatch (the u-values). These are transformed and combined into scores ('match weights') corresponding to the likelihood that a record pair is a match. Decisions about linkage are then based on these scores, usually employing thresholds set by the data linker.

Probabilistic linkage can be more complicated than deterministic linkage because of the additional requirement to estimate m- and u-values (which can be estimated for every matching variable and for every value of every matching variable). It also requires more computing and human resources, mainly because of the need to store and examine match weights. The fundamental distinction between deterministic and probabilistic linkage, though, is that the former is based on rules and the latter on weights or scores. The remainder of this article will now be devoted to explaining why this distinction between rules and scores has little real substance after all (don't stop reading!) and neither do many of the other distinctions that are often drawn. There are, however, some nuanced distinctions that do matter and many important design elements that can be implemented in either approach. These will become clearer as we proceed through the myths and will then be summarised at the end. More detail on deterministic and probabilistic linkage methods and terminology can be found elsewhere (7, 10-12).

Myths and misconceptions about probabilistic linkage

Myth: Probabilistic linkage and deterministic linkage are completely distinct methods

Deterministic linkage classifies record pairs according to decision rules relating to agreement over a set of matching variables. With multiple matching variables, there can be a great many potential patterns of agreement. For example, 'agree-agree-disagree' would be one of eight (2^3) possible patterns for binary agreement/disagreement on three identifiers; but for ten binary measures of agreement, there are $2^{10} = 1024$ possible patterns of agreement. Allowing for partial agreement and missing identifiers significantly increases the number of potential agreement patterns. The task of identifying which decision rules to use in a deterministic procedure therefore becomes increasingly difficult as the number of matching variables increases. Probabilistic linkage is essentially a way of using statistical theory to inform the selection of decision rules. By associating each pattern of agreement with a score, the match weights generated in probabilistic linkage provide a basis for ranking all of the possible decision rules. Once a threshold has been chosen with which to classify probabilistic match weights, the procedure becomes essentially deterministic; record pairs with match weights exceeding the threshold are classified as links and those below the threshold as non-links.

Truth: Every possible decision rule that could be specified in deterministic linkage corresponds to a match weight in probabilistic linkage. Every possible match weight threshold that could be specified in probabilistic linkage corresponds to a set of decision rules that could have been specified in deterministic linkage. They are fundamentally equivalent; differences arise in practice because they are implemented in different ways.

Myth: Probabilistic linkage is based on the probability that record pairs are a match

Arguably 'the one myth that binds them all', this is probably also the most difficult to appreciate. This myth is frequently propagated in the introductions of journal articles; even our own may have misled you. The explanation for this myth may seem nit-picky, but it is nevertheless important for understanding the distinction—or lack thereof—between probabilistic and deterministic linkage.

In theory, match weights vary according to the likelihood that the pair of records belong to the same person. Here lies the crux of the matter: match weights are not likelihoods but rather a score that, if certain assumptions hold, gets bigger when the likelihood of being a match is high, and smaller when it is low. The likelihoods of each pattern being a match are rarely known, and the standard assumption used to generate match weights (that m- and u-values are statistically independent) may not hold, meaning that match weights do not always correspond with the likelihood that the pair is a

match (for further explanation see Winkler (2) and for some alternatives that allow for dependence between identifiers, see Goldstein, Harron (13) or Daggy, Xu (14)).

Truth: Match weights are scores that are expected to correlate with the likelihood that a record pair is a match given the observed pattern of agreement. Match weights are not probabilities, though, even when rescaled to fit in the 0-1 probability space. The likelihood that a pair is a match is neither known nor estimated.

Myth: Probabilistic linkage is *intrinsically* imperfect or imprecise

Because match weights are not actually probabilities, the name probabilistic linkage is itself misleading. The belief that probabilistic linkage is based on probabilities seems to lead some people to infer that there is an inherent degree of error involved (because probabilities are by definition less than perfect). We're not sure when the term *probabilistic* first started being applied to Fellegi and Sunter's method (Jaro (15) at least helped the term gain traction), but they use it themselves only in a few specific contexts and not as a general label for the theory. While the label is not inaccurate (match weights should at least correlate with likelihood of a pair being a match and the m- and u-values really are probabilities), we suspect that it has led to a lot of confusion about how the approach actually works. Perhaps there would be less confusion if it were called 'probability scoring'.

Once you appreciate that match weights are not probabilities, and may only be loosely correlated with them, it is relatively easy to imagine situations in which probabilistic linkage could be implemented with zero error. There is nothing to prevent probabilistic techniques being applied with a unique, error-free identifier (match weights would all be either infinite or negative-infinite, but that is not a problem for discrimination). The real value of probabilistic linkage, though, is in extracting the combined discriminatory power from multiple, less-than-perfect or downright crude identifiers-those which are poorly discriminating, such as gender, are not constant or contain large amounts of error. Provided that the identifier inputs have sufficient *combined* discriminatory power, then probabilistic linkage can produce match weights that allow perfect discrimination.

To understand combined discriminatory power, consider three matching variables: name, date of birth and address. None of these is likely to uniquely identify individuals but the combination of all three is very likely to. Variables that have fewer categories, such as gender, are less discriminatory but can still provide some value when combined with other matching variables. It might be relatively easy to specify deterministic rules based on three matching variables but gets more complicated with larger numbers of variables. Harron, Gilbert (16) demonstrate an initial deterministic linkage using seven matching variables, supplemented by a probabilistic linkage using 23. Most of the additional 16 matching variables contained clinical information with small numbers of categories (e.g. delivery method and gestational age at birth) so were individually poorly discriminating. Combining these variables with the initial seven, however, increased the match rate from 42% to 97% (the match rate is the proportion of records linked, which is only useful when you know how many should; in this case 100% of babies were expected to have a mother).

The combined discriminatory power of matching variables is visualised when data linkers plot the distribution of match weights; high quality matching variables produce well-separated bimodal distributions that support good and feasibly perfect discrimination of matches and non-matches; fewer, poorer quality matching variables produce poorly separated distributions that indicate higher degrees of error at any threshold (Figure 1). Just as with deterministic linkage, how well probabilistic techniques perform is primarily determined by the quality of the matching variables.

Truth: It is possible for probabilistic match weights to facilitate perfect discrimination of matches and non-matches. In both deterministic and probabilistic linkage, whether or not perfect discrimination is possible depends primarily on quality of the matching variables—their uniqueness, errors and missing data.

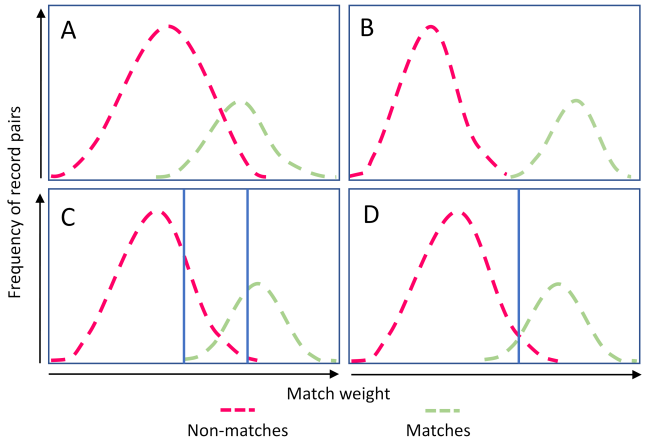
Myth: Probabilistic linkage produces more false matches | Deterministic linkage produces more missed matches

Record linkage, like medical diagnosis, is a classification problem; the aim is to classify record pairs as either matches or non-matches. Just as there is always a trade-off between false positives and false negatives in any diagnostic screening test, there is also an inescapable trade-off between missed matches and false matches. Reflecting this similarity, the proportion of true matches linked is called the *sensitivity* of linkage and the proportion of non-links that are correctly classified as such is the *specificity*.

The principle output of deterministic linkage is a binary variable, 'meets rule/does not meet rule', or sometimes a 'step' or 'match rank' that indicates which rule was met in a series of rules. The principle output of probabilistic linkage is a score, or match weight. Whether deterministic linkage produces more missed matches or false matches depends on which decision rules were selected; in probabilistic linkage, it depends on where the decision thresholds are set (and potentially how badly the independence assumption is violated, although there is conflicting evidence about this and it is an ongoing topic of research (13, 17, 18)). Since each decision rule can be associated with a match weight, selecting deterministic rules and setting probabilistic thresholds are fundamentally the same problem, . It is possible to select only decision rules that favour avoiding false matches (e.g. by incorporating larger numbers of matching variables or only highly unique ones), or to include some which favour capturing more true matches (using fewer or less unique matching variables), just as it is possible to set the probabilistic threshold higher or lower. In most cases, for any probabilistic threshold that can be specified, there is an equivalent set of deterministic rules that could have been specified instead.

There are, however, some kernels of truth to this myth; ways in which probabilistic linkage can perform better than deterministic linkage at least in terms of sensitivity and potentially in terms of specificity too. The foremost of these is probably that it is easier to incorporate more matching variables in probabilistic linkage. More matching variables means more power to discriminate between matches and nonmatches

Figure 1: Example distributions of match weights and thresholds



Curves illustrating the expected distribution of probabilistic match weights for matches (green) and non-matches (red). In practice, only a single distribution is visualised, representing the match weights for all pairwise comparisons. Classification of links (assumed matches) then generally involves the specification of thresholds (blue). A: Poor discrimination between matches and non-matches (high potential for linkage error); B: Good discrimination between matches and non-matches (low potential for linkage error); C: Two thresholds with manual review region (potential errors subject to review); D: Single threshold and no manual review region (linkage errors accepted).

(e.g. plot B rather than A in Figure 1). There are also at least two ways of measuring agreement on a given matching variable that can generally only be implemented in probabilistic linkage: distance measures, and frequency-based weights.

Distance measures record the degree of partial agreement on an identifier, such as the number of characters that differ or the time between two dates. By accommodating errors or inconsistencies in identifiers, distance measures can improve sensitivity of linkage. There are deterministic procedures for accommodating partial agreement that will be discussed further below, but not for factoring in ordinal or continuous measures of agreement.

Frequency-based weights are an option in probabilistic linkage that allows for higher scores to be given to agreement on rarer values. For example, agreement on the surname 'Smith' might be given a lower weight than agreement on 'Doidge'. While it might be technically feasible to incorporate some type of frequency specification into a set of deterministic rules, it would be prohibitively complex for any high-dimensional identifiers such as names.

While these features of probabilistic linkage can give it an edge over deterministic linkage, the reason for this myth's existence is more likely that they are often implemented in different settings, with different priorities for minimisation of missed and false links. When linkage is being conducted for service provision, false links can be especially harmful (e.g. mixing up people's medical or criminal histories). In these settings, highly specific deterministic rules are often sufficient for minimising false links and they provide a high degree of transparency about the minimum criteria for accepting a link. Conversely, in research settings, some false links may be acceptable if it means capturing a great many more missed links; as long as there are estimates of the rates of linkage error, then it may be possible to adjust for it in analysis of the data. Because of this, deterministic and probabilistic linkage are often implemented in different settings (service provision vs. research) and so are often *designed* with different preferences for the trade-off between false matches and missed matches.

Truth: There is always a trade-off between missed links and false links. Both procedures can be tailored towards minimisation of one or the other, or to strike a balance between the two. There are, however, features of probabilistic linkage that can allow it to perform better than deterministic linkage in terms of both missed links and false links, in certain applications.

Myth: Probabilistic linkage requires manual (clerical) review

One commonly stated reason for not conducting probabilistic linkage is a lack of resources for conducting manual review. Manual review is a mechanism often employed to improve probabilistic linkage by subjecting uncertain links to an additional layer of human assessment. Rather than using one threshold to classify pairs as links or non links, two thresholds are chosen. Records with weights above the upper threshold are classified as certain links, records with weights below a lower threshold are classified as certain non-links, and those with weights falling between the two are subjected to manual review (Figure 1C). Thus, the amount of manual review depends on how the thresholds are chosen: the closer the thresholds, the smaller the manual review region. If a single threshold is chosen, no records are classified as uncertain, and no manual review is required. It would be technically feasible to incorporate manual review into a deterministic procedure too, and in practice this is often done as a means of assessing

the sensitivity (true match rate) and specificity (false match rate) of potential decision rules.

Choice of thresholds is not a straightforward matter, yet is no more problematic than deciding on a set of decision rules in deterministic linkage. Choosing a threshold in probabilistic linkage is equivalent to selecting decisions rules in deterministic linkage; a decision needs to be made about which records we are prepared to accept as matches and which we are willing to discard. The benefit of probabilistic linkage is that it provides a ranking of possible decision rules (based on match weights) that is driven by data and explicit statistical assumptions, to guide how we choose to classify records.

Truth: Manual review is an optional component in either probabilistic or deterministic linkage.

Myth: Probabilistic linkage allows for disagreement on matching variables | Deterministic linkage does not

Aside from the one pattern of 'agrees on all matching variables', every other agreement pattern/potential decision rule involves some degree of disagreement. Deterministic linkage approaches often involve setting more than one decision rule and testing them in series until one is satisfied or all are not. Allowance for disagreement in deterministic decision rules can either be implicit, by excluding certain matching variables, or explicit, involving conditions like 'does not disagree on more than one matching variable'. Even when only a single decision rule is used that requires full agreement on a set of matching variables (or a single unique identifier), there are usually other potential matching variables that could have been included but are instead ignored, with implicit allowance for disagreement on these.

Truth: Both probabilistic and deterministic linkage can allow for disagreement on matching variables and almost always do, either explicitly or implicitly.

Myth: Probabilistic linkage can accommodate partial agreement | Deterministic linkage cannot.

Partial agreement occurs when, for example, records match on month and year of birth but not on day of birth, or when text-based matching variables like names are similar (John vs Jon). The date example hints at a strategy that is commonly employed for accommodating partial agreement: breaking a single identifier (e.g. date) down into component parts (day, month and year). Soundex and other phonetic algorithms are ways of accommodating limited degrees of disagreement on text-based matching variables, as are approaches like discarding all but the first three letters of a name. These techniques are all regularly implemented with both probabilistic and deterministic procedures. While not common practice, it would also be technically feasible for a deterministic rule to specify a similarity threshold for agreement on a single identifier, such as 'at least 90% similarity' or 'no more than one character different'.

Truth: Both probabilistic and deterministic linkage can accommodate partial agreement but probabilistic linkage provides a more flexible way of accommodating distance/similarity measures.

Myth: Probabilistic linkage reflects uncertainty in linkage | Deterministic linkage does not

A truly probabilistic *analysis* of linked data would indeed account for uncertainty in linkage, and it is theoretically possible for both probabilistic and deterministic linkage to support this (19, 20). In practice, however, linked data are almost uniformly analysed deterministically, regardless of whether the linkage was conducted using probabilistic or deterministic techniques. What we mean by this is that in nearly every analysis, links (and the absence of links) are treated as error-free. Occasionally, analysts will implement sensitivity analyses, in which a different probabilistic threshold or deterministic rule-step is adopted. This goes a small way towards fully reflecting uncertainty in linkage in the analysis. To probabilistically reflect the uncertainty in linkage, multiple imputation procedures can be used (19, 21) but these methods are not yet in common practice. There are still many unanswered questions about how to properly account for linkage uncertainty in an analysis.

Truth: There is nearly always uncertainty associated with both probabilistic and deterministic linkage. Techniques available for measuring it or accounting for it analytically are limited but emerging. Both probabilistic match weights and deterministic rule-steps can provide a crude indication of uncertainty in a link.

Hard truths about probabilistic linkage

In the interest of completeness, it is worth acknowledging the reasonable grounds that do exist for not implementing probabilistic linkage. Probabilistic linkage does require more computational resources, which can be a constraint, particularly when dealing with very large numbers of records (although it is worth pointing out that population-level, routine ongoing linkage involving dozens of administrative databases have been implemented using probabilistic linkage procedures (22)). Probabilistic linkage is also more complicated so requires relevant expertise and can involve more person-time to implement. There are also settings in which the standard assumptions used to estimate match weights may not be appropriate (for example, due to dependence between errors in matching variables), but alternatives are available (13, 14).

Summary

Hopefully by now you can appreciate that there really aren't that many *intrinsic* differences between probabilistic and deterministic linkage. Many of the claims made about deterministic linkage do not reflect the variety of ways in which it *could* be implemented. Many of the claims made about probabilistic linkage are based on a misinterpretation of match weights as being true likelihoods. As long as the process for linkage is equally well conceived (i.e. that comparable choices relating to agreement on matching variables are reflected in each),

Summary Box

Myth 'Probabilistic linkage...'	**Truth**
...and deterministic linkage are completely distinct methods.'	Each pattern of agreement over matching variables corresponds to a potential decision rule in deterministic linkage and a match weight in probabilistic linkage. For any match weight threshold that can be set, there is generally an equivalent set of deterministic rules that can be specified.
...is based on the probability that record pairs are a match.	It is based on a *score* that, under certain assumptions, *correlates* with the likelihood that record pairs are a match.
...is *intrinsically* imperfect or imprecise.'	The effectiveness of any linkage procedure depends on the quality of the matching variables. Probabilistic and deterministic linkage can be equivalent when the same matching variables are used but it is easier to incorporate poor-quality matching variables in probabilistic linkage.
...produces more false matches.'	There is always a trade-off between false matches and missed matches. In probabilistic linkage, this trade-off can be tuned in either direction by adjusting the match weight threshold.
...requires manual review.'	The use and amount of manual review depends entirely on how the thresholds are chosen and the degree of certainty acceptable in results. With a single threshold, no manual review is required.
...allows for disagreement on matching variables.'	Deterministic linkage also allows for disagreement on matching variables.
...can accommodate partial agreement.'	Deterministic linkage can also accommodate partial agreement.
...reflects uncertainty in linkage.'	In their usual forms, neither probabilistic nor deterministic linkage account for uncertainty in linkage (this is the task for the analysis, not the linkage). Both Probabilistic match weights and deterministic rule steps are crude indicators of uncertainty in a link.

both methods can produce successful linkage. However, optimising a deterministic linkage algorithm becomes very difficult when the data are complicated by large numbers of matching variables, low predictive powers or lots of errors. It is this capacity of probabilistic linkage to handle more and poorer quality matching variables, combined with distance measures and frequency-based weights that allow it to perform better in many applications.

None of this is to say that there are not significant differences in how probabilistic and deterministic linkages are implemented in practice; deterministic procedures are generally more simplistic and more susceptible to influence by the design choices made by data linkers (i.e. in choosing which rules to use when there are many matching variables available). Probabilistic linkage is arguably more data-driven or empirical, with these choices informed by observations and specified assumptions (although the results of probabilistic linkage do depend on the choice of threshold, which is often a subjective decision). It is also true that deterministic rules as they are usually specified typically favour avoiding false matches over avoiding missed matches, but this is mainly a consequence of the decisions made in specifying those rules, not any intrinsic property of the procedure. Probabilistic linkage is very good at making complex linkage applications more feasible and at improving linkage quality with messy data, but this is because messy data makes selecting and deterministic rules difficult, not because probabilistic linkage *performs* intrinsically better under these conditions.

However, because probabilistic linkage can perform better, reluctance against implementing it generally undermines the quality of the linkage conducted. We propose that the only setting in which deterministic linkage is truly justifiable is when only very small numbers (e.g. <5) of high quality matching variables are available. Even in these settings, we would argue that a data linker could still be better off with a simple probabilistic procedure (i.e. one without substantial manual review or distance measures) than with a deterministic approach. The main exception to this is when computing resources present a constraint; a combined deterministic and probabilistic approach can help with this, by reducing the number of records to be probabilistically linked. Deterministic linkage can also provide a 'gold standard' for deriving m- and u-values of matching variables to be used in a subsequent probabilistic stage, if the pairs who match deterministically are representative of the remainder in terms of those variables. An example of this is provided in (16).

Appreciating that probabilistic and deterministic linkage are fundamentally similar allows us to realise some ways in which the implementation of each can be improved. As hinted at in the final myth, it is possible for both deterministic and probabilistic linkage to support probabilistic *analysis* of the data. Some work has already been done in this area (19) and we are currently developing methods that we hope will push the boundaries of what can be achieved with analysis of linked data.

Acknowledgements

JD is funded by the Economic and Social Research Council (grant reference number ES/L007517/1) establishing the Administrative Data Research Centre for England (ADRC-E). The ADRC-E is led by the University of Southampton and run in collaboration with University College London, the London School of Hygiene and Tropical Medicine, the Institute for Fiscal Studies and the Office for National Statistics (ONS) and supported by the Farr Institute of Health Informatics Research (MRC Grant Nos: London MR/ K006584/1) and the NIHR Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. KH is funded by the Wellcome Trust (103975/Z/14/Z).

The myths and misconceptions presented were collected from various sources, including published literature, conference presentations and informal discussions with colleagues and associates. In some cases, no doubt, they reflected oversight or haste, rather than true misconception. For this reason, we have chosen not to cite specific examples of each myth. We would like to thank Prof Ruth Gilbert and Dr Kerina Jones for their comments on early drafts of this article, Prof Harvey Goldstein for his general guidance about all things statistical, and our peer reviewers for their important contribution.

References

1. Fellegi I, Sunter A. A theory for record linkage. J Am Stat Assoc. 1969;64. http://doi.org/10.1080/01621459.1969.10501049

2. Winkler WE. Overview of record linkage and current research directions. Washington, D.C.: U.S. Census Bureau; 2006. Available from: https://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

3. Clark DE, Hahn DR. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. Proc Annu Symp Comput Appl Med Care. 1995;1995. PMCID: PMC2579122.

4. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol. 2010;64(5):565-72. http://doi.org/10.1016/j.jclinepi.2010.05.008

5. Bradley CJ, Given CW, Luo Z, Roberts C, Copeland G, Virnig BA. Medicaid, Medicare, and the Michigan Tumor Registry: a linkage strategy. Med Decis Making. 2007 Jul-Aug;27(4):352-63. PubMed PMID: 17641138. Epub 2007/07/21. eng http://doi.org/10.1177/0272989x07302129

6. Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med. 2002;21. http://doi.org/10.1002/sim.1147

7. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol. 2016;45(3):954-64. http://doi.org/10.1093/ije/dyv322

8. Winkler WE. Probabilistic linkage. In: Harron K, Goldstein H, Dibben C, editors. Methodological developments in data linkage. Wiley series in probability and statistics. Chichester, UK: John Wiley & Sons, Ltd; 2016. p. 8-35.

9. Harron K, Goldstein H, Dibben C, editors. Methodological developments in data linkage. Chichester, UK: John Wiley & Sons, Ltd.; 2016.

10. Harron K. An introduction to data linkage. Administrative Data Research Network; 2016. Available from: https://adrn.ac.uk/media/1324/datalinkage.pdf.

11. Herzog TH, Scheuren F, Winkler WE. Data quality and record linkage techniques. USA: Springer Verlag; 2007.

12. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang LC, Smith P, et al. GUILD: GUidance for Information about Linking Data setsdagger. Journal of public health (Oxford, England). 2017 Mar 28:1-8. PubMed PMID: 28369581. Epub 2017/04/04. eng http://doi.org/10.1093/pubmed/fdx037

13. Goldstein H, Harron K, Cortina-Borja M. A scaling approach to record linkage. Stat Med. 2017:n/a-n/a. http://doi.org/10.1002/sim.7287

14. Daggy JK, Xu H, Hui SL, Gamache RE, Grannis SJ. A practical approach for incorporating dependence among fields in probabilistic record linkage. BMC Med Inform Decis Mak. 2013;13:97-. PubMed PMID: PMC3766252. http://doi.org/10.1186/1472-6947-13-97

15. Jaro MA. Probabilistic linkage of large public health data files. Stat Med. 1995;14(5-7):491-8. http://doi.org/10.1002/sim.4780140510

16. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. PLoS One. 2016;11(10):e0164667. http://doi.org/10.1371/journal.pone.0164667

17. Tancredi A, Liseo B. A heirarchical Bayesian approach to record linkage and population size problems. The Annals of Applied Statistics. 2011;5(2B):1553-85. http://doi.org/10.1214/10-AOAS447

18. Harron K, Hagger-Johnson G, Gilbert R, Goldstein H. Utilising identifier error variation in linkage of large administrative data sources. BMC Med Res Methodol. 2017 Feb 07;17(1):23. PubMed PMID: 28173759. PMCID: PMC5297137. Epub 2017/02/09. eng http://doi.org/10.1186/s12874-017-0306-8

19. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. Stat Med. 2012;31(28):3481-93. http://doi.org/10.1002/sim.5508

20. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Med Res Methodol. 2014;14(1):36. http://doi.org/10.1186/1471-2288-14-36

21. Lahiri P, Larsen MD. Regression Analysis with Linked Data. Journal of the American Statistical Association. 2005;100(469):222-30. http://doi.org/10.1198/016214504000001277

22. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res. 2012;12. http://doi.org/10.1186/1472-6963-12-480