A SIMPLE SAMPLING METHOD FOR ESTIMATING THE ACCURACY OF LARGE SCALE RECORD LINKAGE PROJECTS

Authors

JH Boyd*[1]

T Guiver[2]

SM Randall[1]

AM Ferrante[1]

JB Semmens[1]

P Anderson[2]

T Dickinson[3]

*Corresponding Author (Tel: +61 8 9266 4986; e-mail: j.boyd@curtin.edu.au)

[1]Centre for Population Health Research

Faculty of Health Sciences

Curtin University

Bentley 6102 WA

Australia

[2]Australian Institute of Health and Welfare

1 Thynne Street

Fern Hill Park

Bruce ACT 2617

Australia

[3]Statistics New Zealand

Statistics House

The Boulevard

Harbour Quays

PO Box 2922

Wellington 6140

New Zealand

SUMMARY

Background

Record linkage techniques allow different data collections to be brought together to provide a wider picture of the health status of individuals. Ensuring high linkage quality is important to guarantee the quality and integrity of research. Current methods for measuring linkage quality typically focus on precision (the proportion of incorrect links), given the difficulty of measuring the proportion of false negatives.

Objectives

The aim of this work is to introduce and evaluate of a sampling based method to estimate both precision and recall following record linkage.

Methods

In the sampling based method, record-pairs from each threshold (including those below the identified cut-off for acceptance) are sampled and clerically reviewed. These results are then applied to the entire set of record-pairs, providing estimates of false positives and false negatives. This method was evaluated on a synthetically generated dataset, where the true match status (which records belonged to the same person) was known.

Results

The sampled estimates of linkage quality were relatively close to actual linkage quality metrics calculated for the whole synthetic dataset. The precision and recall measures for seven reviewers were very consistent with little variation in the clerical assessment results (overall agreement using the Fleiss Kappa statistics was 0.601)

Conclusion

This method presents as a possible means of accurately estimating matching quality and refining linkages in population level linkage studies. The sampling approach is especially important for large project linkages where the number of record pairs produced may be very large often running into millions.

Key words:

Medical Record linkage; Electronic Health Records; Linkage quality; Sampling; Estimation

1. INTRODUCTION

In order to manage, monitor, assess and review a range of services, most government departments invest a significant amount of time and effort into collecting and analysing administrative datasets. These datasets are often used in research to provide insight into social issues, to support policy development and improve service delivery (1, 2).

Secondary use of administrative data collections is enhanced through record linkage. This process allows data from different sources to be brought together to provide richer information. The benefits of linked data include reduced data collection costs and more detailed and extensive analysis (3-7).

Record linkage is the process of bringing together data belonging to the same person from within and across different datasets. The process involves comparing identifying information between records to assess whether they belong to the same individual (8). Where there is no reliable unique identification number on the datasets, the matching comparisons typically involve a variety of rules which are applied to available identifying data fields (e.g. name, address and date of birth).

Undertaking record linkage would be easy if identifying information and personal circumstances did not change and were consistently reported. However this is rarely the case and to find all appropriate records, linkage techniques must allow for data imperfections or changes in personal identifiers over time (9).

Although a number of matching methods are available (10), probabilistic methods are generally considered to be the most flexible and reliable when linking large administrative datasets (11-13). Probabilistic methods are also useful if the data involved contains information on individuals with recording discrepancies across the available data fields for matching (14). Using probabilistic matching, record comparisons are assigned a ‘weight’ or ‘comparison score’ based on the agreement of information between records (15). This process allows some tolerance for differences between records with the comparison score corresponding to the likelihood that two records belong to the same person.

Over the last fifty years, sophisticated linkage methods have been developed to allow reliable matching of administrative data (11-13, 16, 17). The challenge across datasets is always the

same; to optimise linkage techniques to ensure they find all records belonging to the same person.

While the basic processes of determining which records belong to the same person using personal identifiers is well established, ensuring high linkage quality is difficult and typically requires a large amount of effort (9).

In many linkage systems across the world, methods of clerical intervention are used in the record linkage process to both evaluate an overall matching approach and to improve matching quality for specific subgroups (18). The clerical processes required during linkage management usually involve a method of validation to ascertain the impact of the linkage procedures (Clerical Assessment) and manual review of potential matches to confirm links (Clerical Review). These processes typically involve human assessment of record pairs to assess or augment the automated linkage algorithm or make a determination when the algorithm cannot reliably confirm or reject a link.

In large scale application, the number of records to be assessed for full clerical review/assessment may be very large often running into several million records.

Most linkage systems can be tuned to optimise the false positive and false negative rates. However, all research projects are different, some require links that are highly accurate while others emphasis maximising linkage rates. Knowing that linkage error can impact on the interpretation of research findings and introduce bias to research studies highlights a need for routinely measuring linkage quality (19). Although standard methods are available to assess the level of false positive matches produced through linkage, it has not always been easy to accurately estimate the number of missed matches (20).

With an ever increasing number of research studies involving linked data, researchers are requesting information on matching quality to ensure the appropriate analyses can be performed (19, 21). In this work we suggest a simple and replicable approach to address this information deficit.

2. OBJECTIVES

Our primary aim was to introduce and evaluate a sampling method to estimate different aspects of matching quality following record linkage. Developing standard methods to

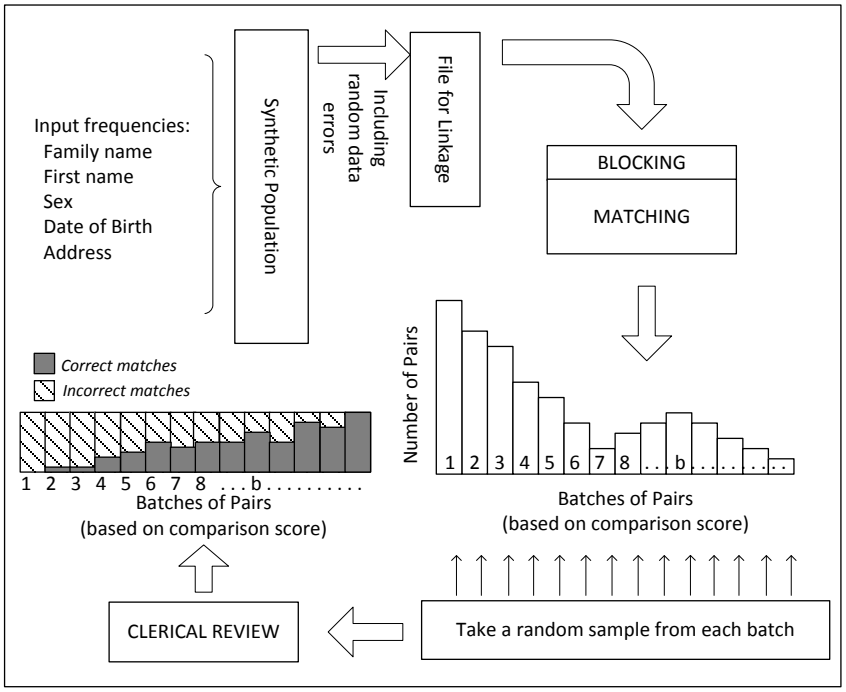
measure matching quality is an important area which can be used by linkage units to refine linkage strategies and inform subsequent analysis; this is essential as the number and complexity of record linkage infrastructure projects continues to expand.

In introducing this method, our first objective was to examine the suitability of the proposed sampling technique, in terms of the accuracy and consistency of estimated linkage quality metrics, for large scale (and enduring) record linkage projects. The second objective was to assess the consistency of clerical review as an assessment tool and to identify potential variation in the process. Finally, a study of inter-rater reliability examined the extent to which two or more individuals agree on the assessment of possible matches (i.e. do reviewers perform relatively similarly) and whether automated assessment procedures can be trained to undertake resource intensive clerical processes (22).

3. METHODS

As complete clerical review of large sets of record pairs is often not feasible, the research team proposed and evaluated a method of estimating linkage quality using sampled clerical assessment. The study used clerical review of the sampled linked pairs to estimate the total false positive and false negative rates at each linkage matching score. Figure 1 outlines the methods used in this paper.

Figure 1 – Sampling technique: flow diagram



3.1. Creation of test data

The evaluation used 'synthetic' datasets to assess the quality of the linkage. 'Synthetic data' is the name given to artificially created records that have characteristics closely resembling the attributes of real world datasets (23). Such datasets are typically used in benchmarking or systems testing (24, 25).

For our purposes, we selected the data generation programme that was developed and implemented as part of the open source FEBRL data linkage system (26). The generator was originally developed in 2005 and is based on ideas by Hernandez and Stolfo (27). It is argued to be an improvement on other generators such as the UIS Database Generator (28) and the generator by Bertolazzi and colleagues (29).

For the study we generated datasets that were suitably representative (i.e. based on real world frequency and error distributions) and of sufficient size to enable realistic testing of both the sampling and linkage quality assessment methods (20). Generation of synthetic data was broken into two stages: (i) creation and use of a large, representative version of the population i.e. a population file, containing 2 million records (1 record per person); and (ii) generation of duplicate records with errors (in our case, synthetic records with repeat events) based on this population i.e. a file for 'linkage' which contained 495,369 simulated events for 47,337 individuals.

Each record in the datasets comprised the following data items: family name, first name, sex, date of birth and address. Records in each dataset were generated with errors typically found in administrative data. Ascertaining representative rates of different types of errors such as duplications, omissions, phonetic alterations and lexical errors involved abstracting errors manually from a number of real world datasets and extrapolating these to the artificial data.

The advantage of the synthetic data over real world datasets is twofold. Firstly, the synthetic data does not have the same strict confidentiality or privacy restrictions in terms of sharing and access. Secondly, the synthetic data was tailored for the project and designed to provide the truth set to evaluate both the sampling and assessment techniques.

3.2. Specification for the Record Linkage

The record linkage process involved a deduplication (internal link) of the synthetic 'linkage' data file aiming to identify all records belonging to each individual from within the file. Probabilistic linkage methods were used to internally link the file, owing to their flexibility and simplicity (8, 30-32). The matching process involved a series of comparisons between two records and a decision as to whether they belong to the same individual. The linkage strategy was implemented on the CUPLE linkage software developed by Curtin University.

The linkage strategy in this study followed a typical approach used to ensure the best quality results, based on a previously published 'default' linkage strategy (20, 33). The linkage strategy applied the general framework of Fellegi and Sunter [14]; a number of standard extensions to this framework are applied, such as approximate string comparators (34). A blocking step was included in the linkage framework and was designed to give matches the best chance of being linked (31). Blocking is an initial linkage step that reduces the number of record comparisons, with matching evaluations only made within clusters of one or more identifying ("blocking") variables (15).

The 'blocking' step limited comparisons to those records which shared a minimum level of identifying information. Two blocks were used (Block 1: soundex of surname and first initial; Block 2: date of birth). All possible comparison variables were compared in each block. String similarity measures were used for all alphabetic variables (name, address and suburb) with exact matches being carried out on all other variables. Day, month and year of birth were all compared separately. Agreement and disagreement weights were estimated by manual evaluation of a small number of pairs.

Each variable comparison resulted in a score based on the specific agreement and disagreement weights for that variable. These scores are summed across the variables to produce a record pair comparison weight. This process results in a set of record pairs each of which have a high probability of belonging to the same individual (35).

3.3. Linkage quality estimation

Clerical review of a sample of record pairs was used to assess the quality of the linkages undertaken in the evaluation. While clerical inspection of a record does not necessarily identify the true match status, it provides the most appropriate judgement given the data fields available on each file.

Clerical reviews were undertaken by seven people with a mixture of linkage experience. From these reviewers, an estimated profile of incorrect and correct links in each batch was created. These results were used to establish a profile of estimated incorrect and correct links that would be accepted at each matching score.

In addition, an automated assessment methodology was developed to replicate the type of rules used by reviewers to determine whether records belong to the same person. The automated assessment applied logical rules to decide if the record pairs are 'links' or 'non links'. These logical rules were held outside the system and modified, removed or added to by clerical operators (36).

The automated assessment methodology used an iterative process allowing clerical reviewers to identify additional rules which could be added to the logic to supplement the already available rules and further automate the clerical review process.

The automated assessment was 'trained' for this study by one of the reviewers using real administrative data. The reviewer's knowledge was added and validated incrementally based on their manual clerical review of pairs. The fully saturated model was then applied to the sample synthetic pairs as an automated clerical reviewer.

3.4. Linkage quality metrics

In assessing the quality of linkage, of primary interest is knowing the number of true matches and non matches identified as links and non-links. True matches and true non-matches are not usually known prior to a linkage. However, as the datasets were synthetically generated, it is possible to flag which records belongs to a specific population record. In this way, it is possible to know all the true matches and non-matches a priori.

Linkage quality was evaluated using pairwise precision and recall. These measures have been previously used in record linkage literature (10, 37).

Precision refers to the proportion of returned links that are true matches. It is sometimes referred to as positive predictive value and is measured as:

$$Precision = \frac{Number\ of\ true\ \ positives}{Number\ of\ true\ positives + Number\ of\ false\ positives}$$

Where a true positive is a pair of correctly linked records, and a false positive is one that is incorrectly or falsely linked. False positives are pairs of records that have been falsely linked (i.e. brought together through linkage but actually belong to different people).

Recall is the proportion of all true matches that have been correctly linked. Recall is also known as sensitivity and is measured as:

$$Recall = \frac{Number\ of\ true\ \ positives}{Number\ of\ true\ positives + Number\ of\ false\ negatives}$$

A false negative is a pair of records that should have been linked because they belong to the same person but were not. False negatives or Type II errors are difficult to estimate in real world situations.

In terms of quality metrics, precision and recall were calculated based on both the Actual and Sampled records pairs.

3.5. Sampling record pairs

A stratified sampling scheme was established to produce reasonable estimates of links and non-links at each comparison weight. The record pairs were divided into batches with similar comparison weights, using the integer part of each pair comparison weight, and the proportions of links and non-links in each batch sample, as determined by clerical review, were recorded. As is typical in a standard clerical review process, reviewers were not blind to the pair comparison weight. Estimates of the numbers of confirmed links and false links in the whole batch were calculated using well established statistical sampling theory (38) (39).

Within each batch (b) a sample of record pairs was randomly selected, then the number of confirmed links as matches in the sample ($n_{b,tp}$) and the links not confirmed as matches in the sample ($n_{b,fp}$) were weighted up to the estimate of total confirmed links using 'number raised estimation'. Put simply the proportions of confirmed links and false links observed in each batch were multiplied by the batch size to create estimates for the whole batch. Number raised estimation is a simple estimation methodology, is unbiased, does not require auxiliary data and the accuracy of the estimates can be simply calculated

The estimate of the total number of confirmed links in a batch is given by:

$$\widehat{N}_{b,tp} = \frac{N_b}{n_b} n_{b,tp} \text{ and } \widehat{N}_{b,fp} = \frac{N_b}{n_b} n_{b,fp} \quad (1)$$

Where:

- $\widehat{N}_{b,tp}$ estimate of total confirmed links that are in batch b
- $\widehat{N}_{b,fp}$ estimate of total links not confirmed as matches in batch b
- N_b total number of links in batch b
- n_b number of links in batch b sample
- $n_{b,tp}$ number of confirmed links in batch b sample
- $n_{b,fp}$ number of false links in batch b sample

As the record pairs were divided into batches with similar comparison weights, batch estimates can be aggregated to estimate the total number of true and false positive for the linkage at each threshold weight.. If a decision rule is applied to accept all record pairs in batches b' and above, batch level estimates for these accepted record pairs can be calculated simply:

$$\widehat{N}_{tp} = \sum_{b \le b'} \widehat{N}_{b,tp} \quad \text{and} \quad \widehat{N}_{fp} = \sum_{b \le b'} \widehat{N}_{b,fp} \quad (2)$$

These estimates reflect the number of correct links (linked records presumed to be matches) and incorrect links (linked records presumed to be non-matches).

Each batch estimate is subject to sampling error measured by the variance of the estimate, which we estimate with

$$\widehat{Var(\widehat{N}_{b,tp})} = {N_b}^2 \left(1 - \frac{n_b}{N_b}\right) \frac{\left(\frac{n_{b,tp}}{n_b}\right)\left(1 - \frac{n_{b,tp}}{n_b}\right)}{n_b - 1} \quad (3)$$

These variances can be added up over batches of accepted record pairs:

$$\widehat{Var(\widehat{N}_{tp})} = \sum_{b \le b'} \widehat{Var(\widehat{N}_{b,tp})} \quad (4)$$

These measures of accuracy on counts can be further used to derive measures of accuracy and confidence intervals on derived quality measures (such as precision and recall). These measures of accuracy can be calculated for different thresholds of accepting record pairs.

3.6. Selecting a sample size

The sample size to be used within each batch can be evaluated using power analysis. This ensures that the sample has a reasonable chance of detecting a significant difference. The evaluation used 31 batches based on the integer part of each pair comparison weight. For this investigation a fixed sample size of 100 record pairs was chosen within each batch. At a significance level of 0.1 (α=0.1), a sample of 100 record pairs has over 75% chance of detecting a significance difference in the estimate of a precision or recall of 0.1 (assuming a null proportion of 0.5).

The sampling performance of the individual batches is a starting point for the sampling strategy. The statistical power in the individual batches need not be very high as the aggregated estimates (derived from estimates given by (3)), are subject to less sampling variability the batch level estimates. An optimal sample design could allocate less sample to batches with very high weights or very low weights, as these batches would tend to be more homogenous and thus less variable. A fixed sample for each batch was selected for two reasons. Firstly, without first undertaking an initial investigation or having other evidence it is difficult to determine an appropriate sample size tailored for each batch in advance. Secondly, the homogenous batches tend to be quickly enumerated.

3.7. Reliability Assessment

The sampled pairs were evaluated by seven different reviewers and quality metrics were calculated for each reviewer. A second investigation looked at whether the quality metrics produced by each reviewer were consistent. In order to assess whether there was any significant difference between the quality estimates produced by the reviewers we use the Fleiss Kappa Statistic (22, 40). This method assessed how similar the reviewers were in classifying the pairs into true matches and true non-matches.

$$K = \frac{\bar{\bar{P}} - \bar{\bar{P}}_e}{1 - \bar{\bar{P}}_e}$$

Where $\bar{\bar{P}}$ is the average proportion of agreement among the reviewers, and $\bar{\bar{P}}_e$ is the proportion of agreement among the reviewers are expected to agree by chance alone.

Complete agreement corresponds to K = 1, and lack of agreement (i.e. purely random coincidences of rates) corresponds to K = 0.

4. RESULTS

As expected, the sampling process accurately and precisely represented the matching results for the entire synthetic population. A comparison showed little difference in the percentage of correct links (true matches) at each matching score between the entire synthetic population and the selected sample. The accuracy of the estimates follows on from the fact the number raised estimator is unbiased. The sample size was designed using power analysis to ensure a reasonable level of precision resulting in reasonable precise estimates within each batch. A smaller sample while still accurate (i.e. unbiased) would have been less precise (i.e. have a higher degree of sampling variability).

Linkage of the synthetic data produced a series of records pairs with a matching score. The sample based methodology was applied to the linked record pairs which were divided into batches based on overall comparison weights. Each batch contained record pairs with a comparison weight within a specified interval. The intervals were of equal width and non-overlapping, ensuring each record pair fell into a single batch. In assessing the sampling, our first aim was to investigate how many true matched and non-matched records were identified or returned in each sampled batch and how this compared to all matches.

Using synthetic data we know which records belong to each individual and as a result, all true matches and non-matches from the linkage. From the ‘true’ and ‘false’ links we calculated linkage quality metrics (precision and recall) for both population and sample.

The population and sample precision measures (Figure 1) show the proportion of returned links that are ‘true’ matches for the given comparison weight. This includes all record-pairs that score above this threshold value. The precision curve runs from 0.88 at a matching score of 14 to 1.0 at highest matching scores (i.e. no false positives).

The recall figures (in Figure 2) show the proportion of all true matches that have been correctly identified. The recall curve runs from 1.0 at a matching score of 14 (i.e. all true matches identified) to 0.65 at highest matching scores.

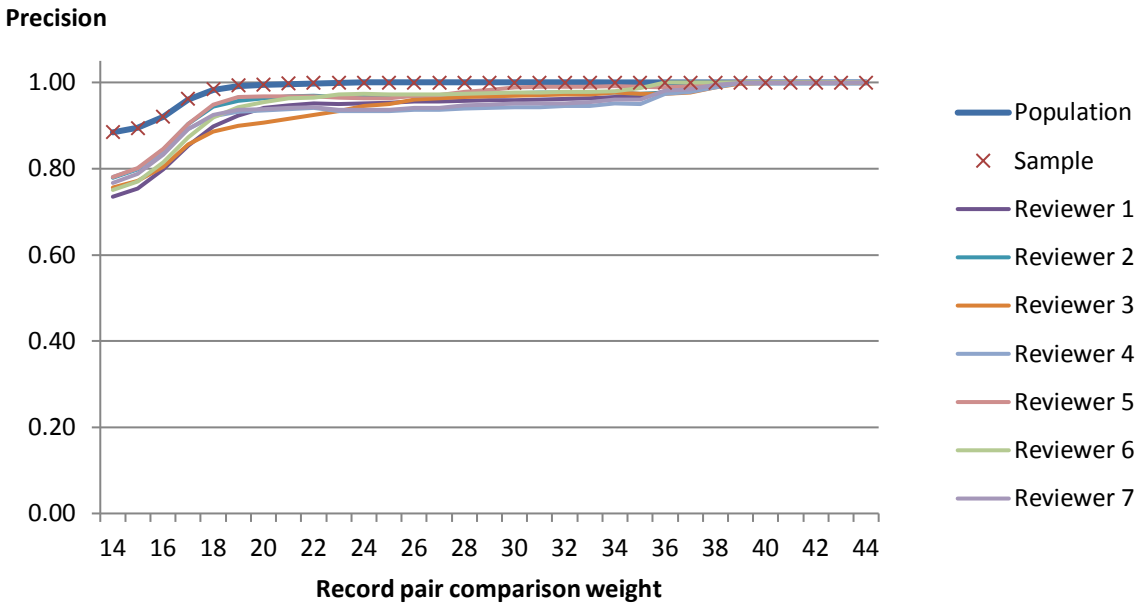
4.1. Reviewer results

The sampled record pairs were clerically reviewed to determine a link status for each (providing an estimate of true positive links and false positive links in each sample batch).

These estimates were used to calculate linkage quality metrics (precision and recall) for each reviewer. The estimated precision and recall results have been presented in Figures 2 and 3.

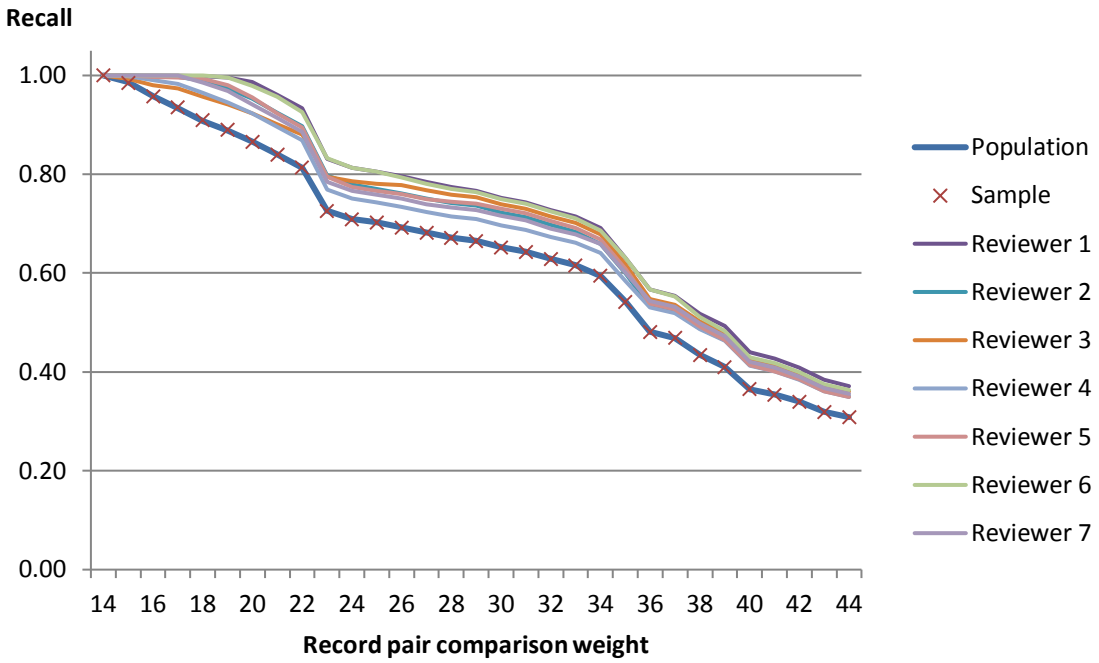
The estimated precision outputs (Figure 2) for all seven reviewers are very consistent with little variation in the clerical assessment results. However, the estimates are slightly lower than the actual results (especially for the lower matching scores) providing a conservative estimate of the true matches.

Figure 2 – Reviewer precision estimates by matching score



The estimated recall outputs (Figure 3) for the seven reviewers are also very consistent. As the number of 'true' matches has been underestimated in the batches with low matching scores, the estimates of recall are slightly higher than the actual results.

Figure 3 – Reviewer recall estimates by matching score

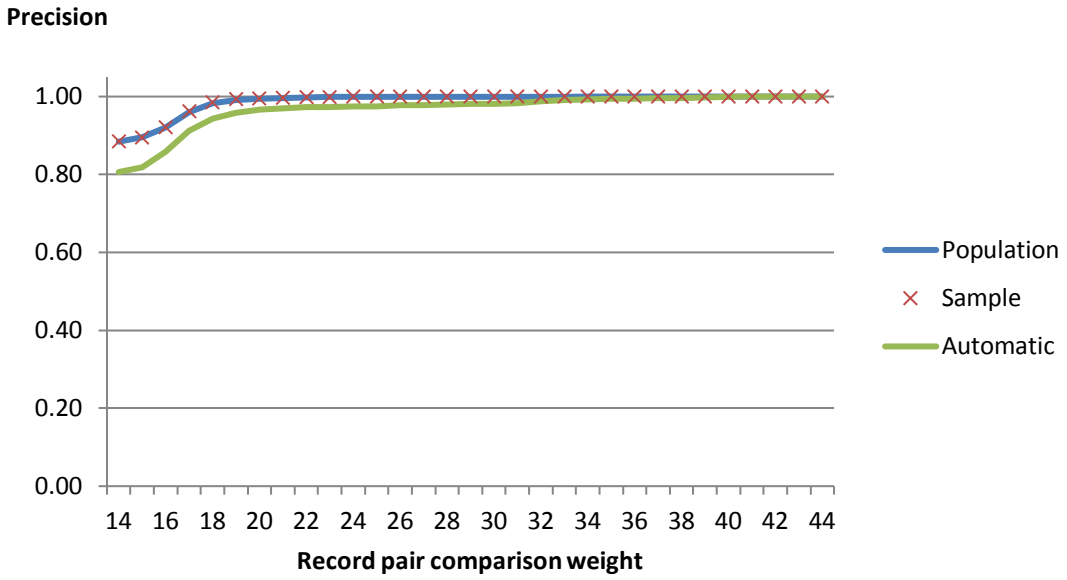


4.2. Automated results

These automated decisions were used to calculate linkage quality metrics. The estimated precision and recall results from the automated tool have been presented in Figure 4 and Figure 5.

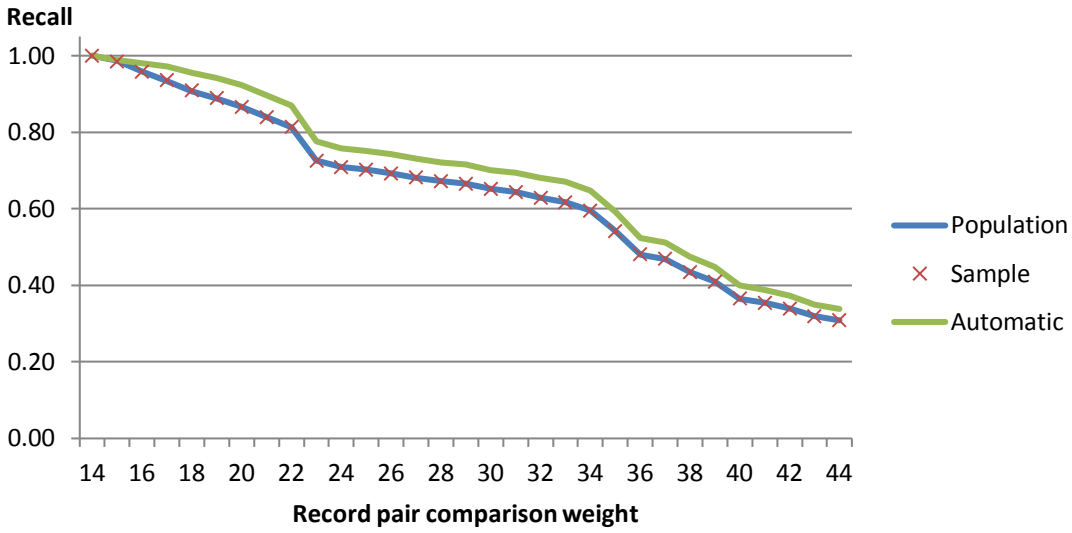
Not surprisingly, as they have been based on the rules from a clerical reviewer, the automated quality estimates are similar to the review results. The precision results are slightly lower than the actual results (especially for the lower matching scores) and the recall is slightly higher. The recall estimates are very stable across the matching score batches.

Figure 4 – Automated review precision estimates by matching score



The results for both precision and recall from the automated tool are close to the actual figures. This suggests that the rules for accepting and rejecting matching pairs are being applied reliably by the automated tool (removing human judgement from the decision).

Figure 5 – Automated review recall estimates by matching score



4.3. Reliability of Assessors Results

Clerical reviews were undertaken by seven people with a mixture of linkage experience. Each provided a profile of incorrect and correct links in each batch. The outputs from each of the clerical reviewers were assessed for consistency using the Fleiss Kappa statistics which measures inter-rater reliability for multiple reviewers/raters.

The overall agreement for the seven clerical reviewers was K = 0.601 (CI 0.593, 0.609). Interpretation of the Kappa statistic generally suggests that results over 0.60 suggest a good strength of agreement between the reviewers/raters (22). When the Kappa statistic was restricted to the most experienced reviewers, it remained in the 0.6 to 0.8 range suggesting good agreement.

5. DISCUSSION

In recent years, significant investment in record linkage infrastructure has occurred internationally, reflecting the strategic value of high quality linked datasets (7, 41-44). Although there has been significant development around scalable linkage units to support population level human research there has been little progress in reporting matching quality within these dedicated systems (45). Knowing that both 'wrong links' and 'missed links' can impact on the interpretation of research findings highlights a need for routinely measuring linkage quality (19).

While it is possible to identify false positives based on the results of a linkage (e.g. using targeted clerical review on linkage output), identifying the missed links is more difficult and often left unknown (46). Common quality assurance reporting implementations which contain estimation of false positive and false negatives usually involve complete review of linkage results, 'gold standard' datasets (used as a benchmark to assess linkage quality) and the application of group based logic mapping (e.g. a group of records belonging to a single person which includes a hospital record with a discharge dead code should also contain the associated death registration). However, these techniques are often constrained by the effort involved or the accuracy of the results.

The objective of this paper therefore was to design and evaluate scalable methods of clerical assessment to allow linkage units to assess the quality or accuracy of matching processes and provide research extracts with estimates of the linkage quality (19).

As expected, the sampling method showed no significant difference between the sample percentage of correct ('true') links at each matching score and the actual percentage across all links. The sampling methodology, which uses a simple yet robust probability based sample

design and random selection of matched record pairs across batches, provided an unbiased and reliable sample of all links generated by the linkage.

In addition, the linkage quality metrics calculated on the selected sample were not significantly different to the actual precision and recall metrics for the whole linkage. The results demonstrate that as well as the sample estimates of precision and recall being unbiased (accurate) they are also precise (that is subject to a low level of sampling variability).

The evaluation of the method itself found that using manual inspection of the sample batches to assess the overall linkage provided an acceptable evaluation of the linkage quality. Although there were small differences between the assessors, generally the strength of agreement was good across all clerical reviewers.

The sampling methodology provides a number of advantages in assessing linkage quality. It offers a manageable and cost effective framework for the assessment of linkage quality (and, additionally, threshold setting). By applying this technique it is possible to assess both the accuracy of matches made as part of a linkage and to estimate the proportion of missed links. The assessment of missed links is traditionally difficult to undertake but can be important to researchers who wish to adjust research results based on the overall linkage quality.

Furthermore, in comparison to traditionally expensive processes of clerical assessment, the sampling methodology offers an objective method of quality assessment for probabilistic record linkage without a substantial investment in clerical processes.

This method can also be applied to 'deterministic' record linkage, where instead of the probabilistic approach, a series of logical rules are used to determine which records belong to an individual. In the rules-based approach, rules would need to be ordered based on how 'strict' they were – i.e. the likelihood of containing a false positive. Additional rules would also have to be developed, of a lower quality than those currently used, in order to estimate missed matches.

Another finding in the evaluation was that by using systematic assessment methods to automate the review process it was possible to capture and apply clerical knowledge. The automated reviewing tool provided as good an estimate of both precision and recall as the human reviewers.

While human decision making based on record pairs is the traditional method for quality assessment, there are automated options capable of reducing the manual workload. Although time is required by an assessor to build the knowledge base for the clerical process; this logic, which is added incrementally and identifies any conflicting rules, can be used effectively to refine the matching strategy or enhance comparison routines.

Although the estimated precision and recall results indicate a high level of consistency between the reviewers (including the automated assessment), the estimates are somewhat different to the actual results, providing a conservative estimate of precision and an overestimate of recall (especially for the lower matching scores). Interestingly, if these were combined through the derivation of an F-measure (harmonic mean of precision and recall) the overall affect is reduced.(10) Further testing on additional datasets is required to determine whether this is a systematic, or dataset dependent effect. Feedback from reviewers indicated that the limited matching fields in the synthetic data (without any additional information) and compounded error modifications made identification of 'true' positive links more difficult in batches with a low matching score.

The sample methodology overcomes some of the challenges that have been experienced in estimating the quality of linkage on a manageable amount of clerical review. In general, the method provided acceptable estimates of linkage quality using the synthetic data. The advantage with this over current methods is that will provide an estimate of the overall linkage quality (including missed links). By developing and applying scalable methods of clerical assessment, linkage units can assess the quality or accuracy of the matching process and provide research extracts with the appropriate level of linkage quality (19).

5.1. Limitations

The linkage quality estimation methodology has been specifically designed around the probabilistic record linkage techniques used by many dedicated linkage units. With continuous development in the field of record linkage and scientific progression around matching methods, there should be some consideration of how the approach can be modified to work with any new developments in record linkage algorithms. Appropriate record linkage techniques are often dependent on the quantity and quality of data available and the research context. For clerical review to provide accurate results, reviewers must be aware of all these factors.

The methodology relies on the assessment of matching pairs by reviewers and is often based on subjective judgment to make a decision whether two records belong to the same person. In some circumstances, the clerical reviewer will have more information than used in the linkage strategy upon which they can make an informed decision about whether two records belong to the same person (depending on the data collections). In practice however, the reviewers are often asked to make a decision on the same information used in the linkage process. These decisions can be based on expert knowledge of the dataset but are frequently based on the instinct of the reviewer.

One method which could be explored to improve the manual review would be to modify the method to allow clerical assessment of all pairs belonging to an individual following linkage. Introducing this group checking approach may provide additional information over time to help the assessment process.

While this evaluation has been performed on a large synthetic dataset based on real world characteristics, a more comprehensive analysis could include a variety of administrative datasets. This would provide a wide-ranging assessment using data with different standards and definitions.

6. CONCLUSION

This paper has presented an approach to estimating linkage quality for large scale linkage projects. Our approach provides reliable estimates of linkage quality without full clerical assessment of linkage results. Unlike most estimates, which focus on the accuracy of matches made, this methodology includes missed matches in the calculation of overall linkage quality. Application of the methodology in linkage projects should assist in assessing the performance of linkage operations, customising strategies for specific linkage projects and in the decision making regarding choice of threshold.

7. ACKNOWLEDGEMENTS

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy and Super Science Initiative's Population Health Research Network.

8. REFERENCES

1. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. Health services research. 2010;45(5p2):1468-88.
2. Virnig BA, McBean M. Administrative data for public health surveillance and planning. Annual Review of Public Health. 2001;22(1):213-30.
3. Goldacre M, editor The value of linked data for policy development, strategic planning, clinical practice and public health: An international perspective. Symposium on Health Data Linkage; 2003: Public Health Information Development Unit, Adelaide University.
4. Brook EL, Rosman DL, Holman CDAJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. Australian and New Zealand Journal of Public Health. 2008;32(1):19-23.
5. Hall SE, Holman CDAJ, Finn J, Semmens JB. Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records. International Journal for Quality in Health Care. 2005;17(5):415-20.
6. Sibthorpe B, Kliewer E, Smith L. Record linkage in Australian epidemiological research: Health benefits, privacy safeguards and future potential. ANZ Journal of Public Health. 1995;19.
7. 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 Services Research. 2012;12.
8. Newcombe HB. Handbook for Record Linkage: Methods for Health and Statistical Studies, Administration and Business. New York: Oxford University Press; 1988. 210 p.
9. Boyd JH, Randall SM, Ferrante AM, Bauer JK, Brown AP, Semmens JB. Technical challenges of providing record linkage services for research. BMC medical informatics and decision making. 2014;14(1):23.
10. Christen P, Goiser K. Quality and Complexity Measures for Data Linkage and Deduplication. In: Guillet F, Hamilton H, editors. Quality Measures in Data Mining Studies in Computational Intelligence 43: Springer; 2007. p. 127-51.
11. Roos LL, Wajda A. Record Linkage Strategies: Part 1: Estimating Information and Evaluating Approaches. Winnipeg: University of Manitoba, Medicine Fo; 1990 6 June 1990. Report No.
12. Kendrick SW, Clarke JA. The Scottish Medical Record Linkage System. Health Bulletin (Edinburgh). 1979;51:72-9.
13. Gill LE. OX-LINK: The Oxford Medical Record Linkage System. Record Linkage Techniques. Oxford: University of Oxford; 1997. p. 19.
14. Newcombe H, Kennedy J. Record linkage: making maximum use of the discriminating power of identifying information. . Commun ACM. 1962;5(11):563-6.
15. Fellegi I, Sunter A. A Theory for Record Linkage. Journal of the American Statistical Association. 1969;64:1183-210.
16. Holman D, Bass A, Rouse I, Hobbs M. Population-based linkage of health records in Western Australia: Development of a health services research linked database. Australian and New Zealand Journal of Public Health. 1999;23.
17. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation BMC Health Services Research 2009. 2009;9(157).
18. Rosman D, Garfield C, Fuller S, Stoney A, Owen T, Gawthorne G, editors. Measuring data and link quality in a dynamic multi-set linkage system. Symposium on Health Data Linkage (http://wwwadelaideeduau/phidu/publications/pdf/1999-2004/symposium-proceedings-2003/rosman_apdf); 2002 20-21 March 2002; Sydney.
19. Harron K, Wade A, Muller-Pebody B, Goldstein H, Gilbert R. Opening the black box of record linkage. Journal of epidemiology and community health. 2012;66(12):1198-.

20. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics. 2012;45(1):165-72.
21. Neter J, Maynes ES, Ramanathan R. The effect of mismatching on the measurement of response errors. Journal of the American Statistical Association. 1965;60(312):1005-27.
22. Altman DG. Practical statistics for medical research: CRC Press; 1990.
23. Pudjijono A. Probabilistic Data Generation. Canberra: Australian National University; 2008.
24. Christen P, editor Probabilistic Data Generation for Deduplication and Data Linkage. Sixth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'05); 2005; Brisbane.
25. Nasseh D, Stausberg J. Evaluation of a Binary Semi-supervised Classification Technique for Probabilistic Record Linkage. Methods of information in medicine. 2016.
26. Christen P, editor Febrl - A Freely Available Record Linkage System with a Graphical User Interface. Second Australasian workshop on Health data and knowledge management 2008; Wollongong, NSW.
27. Hernandez MA, Stolfo SJ, editors. The Merge/Purge Problem for Large Databases. Proceedings of the ACM SIGMOD conference; 1995; San Jose, California: ACM New York.
28. Hernandez M. UIS Database Generator. 1997.
29. Bertolazzi P, Santis LD, Scannapieco M, editors. Automated record matching in cooperative information systems. Proceedings of the international workshop on data quality in cooperative information systems; 2003; Siena, Italy.
30. Jaro MA. Probabilistic Linkage of Large Public Health Data Files. Statistics in Medicine. 1995;14:491-8.
31. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association. 1989;84(406):414-20.
32. Copas JB, Hilton FJ. Record Linkage: Statistical Models for Matching Computer Records. Journal of the Royal Statistical Society. 1990;153(3):287-320.
33. Randall S, Ferrante A, Boyd J, Semmens J. The effect of data cleaning on data lnkage quality. BMC Medical Informatics and Decision Making. 2013;13(64):e1.
34. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. 1990.
35. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques: Springer; 2007.
36. Richards D, Chellen V, Compton P, editors. The reuse of ripple down rule knowledge bases: Using machine learning to remove repetition. Proceedings of the 2nd Pacific Knowledge Acquisition Workshop (PKAW'96), Coogee, Australia; 1996: Citeseer.
37. Bishop G, Khoo J. Methodology of Evaluating the Quality of Probabilistic Linking. Canberra: Australian Bureau of Statistics, Analytical Services Branch, 2007 5 April 2007. Report No.: 1351.0.55.018.
38. Cochran WG. Sampling techniques. 1977. New York: John Wiley and Sons.
39. Guiver T. Sampling-Based Clerical Review Methods in Probabilistic Linking. ABS Website (http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1351.0.55.034May%202011?OpenDocument): Australian Bureau of Statistics, Branch AS; 2011 Contract No.: ABS Cat. no.1351.0.55.034.
40. Freelon DG. ReCal: Intercoder reliability calculation as a web service. International Journal of Internet Science. 2010;5(1):20-33.
41. Gill LE. OX-LINK: The Oxford Medical Record Linkage System. Oxford: University of Oxford.
42. Kendrick S, Clarke J. The Scottish Record Linkage System. Health bulletin. 1993;51(2):72.
43. Ford DV, Jones KH, Verplancke J-p, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. 4 September 2009.
44. Roos LL, Nicol JP. A research registry: uses, development, and accuracy. Journal of clinical epidemiology. 1999;52(1):39-47.

45. Lyons RA, Hutchings H, Rodgers SE, Hyatt MA, Demmler J, Gabbe BJ, et al. Development and use of a privacy-protecting total population record linkage system to support observational, interventional, and policy relevant research. The Lancet. 2012;380, Supplement 3(0):S6.
46. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. 2010.