2. Data Linkage in Manitoba

The internationally recognized Manitoba Population Research Data Repository is made possible through the integration of data across multiple domains, including health, education, social services, judicial affairs, and immigration.[6] This extensive collection of routinely-collected data with excellent population coverage of Manitoba residents, enables high-quality interdisciplinary research of topics highly relevant to Canadians, such as the social determinants of health and healthcare delivery patterns.

Databases contained within the Repository are linked using a method known as "spine linkage." In this approach, each dataset is linked to a central "spine dataset" (i.e., the Manitoba Health Insurance Registry), rather than performing pairwise linkages separately between all datasets.[7] After separately linking each Repository database to the Manitoba Health Insurance Registry to attach an encrypted Personal Health Identification Number (PHIN) field to each database, the encrypted PHIN is then used as the common unique identifier by researchers to join the de-identified datasets.[6] Spine linkage makes large data repositories feasible by significantly reducing the number of linkages required, which in turn reduces the disclosure of personally identifiable information.[7] However, a key limitation of spine linkage is the exclusion of individuals who are present in some data sources but not in the spine dataset. Consequently, the practice of performing spine linkage with the Manitoba Health Insurance Registry may adversely impact research based solely on non-health data sources. Studies focused on population subgroups without coverage through the Manitoba Health Services Insurance Plan (e.g., federally insured individuals, foreign students, temporary foreign workers with a work permit duration of less than one year) or with delayed coverage (e.g., new residents from other Canadian provinces) should carefully consider these limitations and changes in coverage requirements over time.[8,9]

In Manitoba, the responsibility for linking research data is shared among MCHP, Manitoba Health, and CHI. MCHP handles linkage between Repository databases and the Manitoba Health Insurance Registry, except for certain federally regulated data sources, which are managed by Manitoba Health. CHI and Manitoba Health share responsibilities for project-specific linkage services, such as the linking of clinical or survey datasets to the Manitoba Health Insurance Registry, enabling researchers to create enriched datasets by combining primary and routinely collected data. Additionally, CHI provides methodological guidance on data linkage practices and offers consultations to researchers and trainees on adjusting for linkage error in data analyses.

To minimize the exposure of personal information in accordance with data privacy legislation, including the Personal Health Information Act (PHIA) and the Freedom of Information and Protection of Privacy Act (FIPPA),[10,11] data linkage using identifiable information is performed by a small, trusted group of analysts from CHI, MCHP, and Manitoba Health. Following linkage, data is de-identified—by encrypting PHINs and removing names and addresses—before data is accessible to researchers within the MCHP secure data analytics environment.[6] Record similarity is commonly assessed using PHIN, given names, surnames, sex/gender, birthdate, and residential or mailing address including postal code. Even when PHIN is present in both data sources being linked, records are never joined solely on PHIN to account for potential data recording errors (e.g., incorrect PHINs being recorded) and to validate the PHINs that were collected.

3. Background on Record Linkage

Record linkage, also known as data linkage and entity resolution, is the task of determining which data records correspond to the same entity (i.e., the same person), and is used to deduplicate data by self-joining records within a single database and to create comprehensive datasets by bringing together person-level information from multiple sources. Record linkage techniques estimate the similarity between all possible pairs of records (i.e., the Cartesian product) by comparing the common fields present in both data sources. The similarity scores are then used to classify candidate record pairs as either matches (i.e., records belonging to the same individual) or non-matches (i.e., records belonging to different individuals). Given the potential for misclassification, the quality of linked data should be carefully assessed before proceeding with research data analysis.

The objective of record linkage is to maximize linkage rate, the proportion of records in the left dataset linking to records in the right dataset, while minimizing false positive matches, which occur when records belonging to different individuals are incorrectly linked. Linkage algorithms commonly consist of multiple stages, with different techniques and parameter combinations selected in each pass to increase the amount of successfully matched record pairs. Identifying fields are combined differently across iterations to minimize data quality issues in select fields from adversely impacting overall linkage performance. Concluding each pass of a multi-step linkage algorithm, all predicted matches are removed from consideration prior to the next iteration.

3.1 Blocking

Examining the similarity of all possible candidate record pairs is computationally infeasible with “big data,” defined as datasets too large to fit into computer memory. For instance, linking two datasets, each containing one million records, results in one trillion candidate record pairs—an impractical number to classify due to excessive computation time and memory requirements.[12] Blocking is a technique used to improve computational feasibility by reducing the search space of candidate record pairs. By constraining comparisons to candidate pairs with exact agreement on one or more identifiers, known as blocking keys, the most dissimilar pairs are filtered out and computational demand is significantly reduced. Blocking is a technique used to improve computational feasibility by reducing the search space of candidate record pairs. By constraining comparisons to candidate pairs with exact agreement on one or more identifiers, known as blocking keys, the most dissimilar pairs are filtered out and computational demand is significantly reduced. For example, using birth year as a blocking key means, only record pairs sharing the same birth year are considered, while all other potential matches are disregarded in that pass of the linkage algorithm. Although blocking aims to reduce the number of clear non-matches under examination, true matches may inadvertently be missed. To address this, multiple linkage passes can be performed with different blocking keys used at each stage. For example, after a pass with birth year as the blocking key, a subsequent pass might use birth month to capture pairs where the birth year was incorrectly recorded in either the left or right datasets. Blocking criteria is progressively less restrictive in subsequent passes since computational demand decreases as linked pairs from previous steps are set aside and removed from consideration.

3.2 Record Linkage Approaches

There are two main linking approaches to record linkage: deterministic and probabilistic. With deterministic linkage, record pairs are classified as a match only if there is exact agreement among all considered fields. If any single identifier disagrees, even slightly, record pairs will remain unlinked. Secondly, deterministic matching assigns equal weight to all identifiers, ignoring differences in discriminative power, which is the effectiveness of a particular field to differentiate between matches and non-matches. For example, when deterministically linking two records based on surname, birth year, and sex, both records must exactly match on all three fields to be considered a match. However, this simplistic approach does not consider that surname has stronger discriminative ability than the other two fields, due to the relatively low frequency of most surnames within a dataset. In contrast, candidate record pairs agree on sex approximately 50% of the time, even among true non-matches, resulting in poor discriminative power. In this example, we may wish to classify record pairs as matches even if they disagree on sex, providing surnames are sufficiently similar. Probabilistic matching is a more robust approach that adjusts for discriminative power and allows record pairs to be classified as matches even when there is disagreement among some of the identifying fields under consideration.

Linkage algorithms commonly consist of one or more deterministic passes, typically yielding the majority of matches in a dataset, before multiple probabilistic linkage steps are used to further increase linkage rate. The more stringent matching criteria inherent in deterministic approaches leads to higher confidence that predicted links represent true matches. In both deterministic and probabilistic approaches, extensions can be incorporated to increase resiliency against data recording errors. Approximate string matching techniques can be incorporated to account for spelling variations in given names (e.g., Meghan and Megan), abbreviations in residential addresses (e.g., Avenue vs. Ave.), and data recording errors such as postal codes entered as R3E 0T8 instead of R3E 0T6. The Jaro-Winkler string similarity measure is often used for this purpose.[13] Additionally, acceptance ranges can increase flexibility when matching numeric fields, such as allowing birthdates to differ by plus or minus one week.

3.3 Probabilistic Linkage

Probabilistic linkage is commonly performed achieved using the Fellegi-Sunter model where record similarity is estimated through the summation of log-likelihood ratios of two conditional probabilities, known as M (for matched) and U (for unmatched) probabilities.[14] Separate log-likelihood ratios, known as partial match weights, are calculated for each identifying field under consideration. The M probability is defined as the probability of field values agreeing, given that record pairs refer to the same entity (i.e., a true match); whereas the U probability is the probability of field values agreeing, given that record pairs refer to different entities (i.e., a true non-match). The M and U probabilities for each identifier may be manually specified, calculated based on observed frequencies in the data being linked, or more commonly, estimated with unsupervised learning using the Expectation-Maximization algorithm.[15] The partial match weights are summed to estimate a total match weight for each candidate record pair meeting blocking constraints. The match weight, which is unbounded, can be converted to a posterior probability with values scaled between 0 and 1.[16] The match weights or posterior probabilities are then used to classify record pairs as matches if they exceed a chosen threshold.

The traditional Fellegi-Sunter approach requires two thresholds to be set: record pairs with match weights below the lower threshold are classified as non-matches, those above the higher threshold are classified as matches, while candidate pairs with weights between these thresholds undergo manual review. However, this conservative approach is largely impracticable when linking large databases, and instead, a single-threshold approach is commonly employed. While a single-threshold approach improves efficiency by omitting the clerical review of potential matches, the uncertainty of the intermediate match weights can lead to increased rates of misclassification. In cases where a record in the left dataset has multiple pairs on the right dataset with similarity scores above the acceptance threshold, the pair with the highest match weight is selected.

3.4 Linkage Error

Merging records from different sources may lead to two types of errors: 1) incorrectly linking records that belong to two different individuals (i.e., false positive matches), and 2) missed matches between records that belong to the same individual (i.e., false negatives). The linkage errors, which may reduce data representativeness, can arise from data entry errors, such as misspelled names; incomplete information, such as missing PHIN; and transient information, such as surname and residential address, that is recorded inconsistently across data sources. Non-random linkage error can lead to selection bias in the resulting linked data. For example, women who change their surname upon marriage may be disproportionately excluded if data capture dates differ considerably between data sources and maiden name is not available in the data. Similarly, individuals with high residential mobility may be disproportionately excluded if residential address is used for linkage but captured inconsistently across databases. Differential linkage error can be assessed by examining linkage rates stratified by sociodemographic characteristics. In situations where a unique personal identifier, such as PHIN, is available in both data sources, it may be used as the ground truth to evaluate classification accuracy.

4. Methods

4.1 Software

All data analytics were performed using R. Data preprocessing was conducted with the datastan package, l was used for record linkage, and the data linkage quality report was generated with linkrep.[17,18]

4.2 Data Pre-processing

The two datasets being linked were first standardized to facilitate accurate and efficient linkage. Data standardization involved renaming variables to standardized names, reformatting fields to consistent data types, cleaning string variables, and recategorizing qualitative variables.

All punctuation was removed from non-numeric fields, and characters were converted to a common case. To reduce the impact of misspellings and variations in names across data sources, given names with common nicknames and alternative spellings were standardized. For example, "Bill" was converted to "William," while "Haley" and "Hailee" were both standardized to "Hailey." These standardized names were then used in select steps of the linkage algorithm to improve linkage rate. Similarly, to reduce variation among residential and mailing addresses, common abbreviations were substituted with their expanded forms. For instance, "Rd" was replaced with "Road" and "St" was standardized to "Street." Fields containing multiple attributes, such as name fields with both primary (i.e., first name) and secondary (i.e., middle name) given names, were split into separate fields. This step was performed to reduce missed matches, particularly among individuals with multiple given and family names, and for those from cultures where the surname precedes the given names.

Categorical variables, such as sex/gender, were standardized to a common set of values across both datasets. For example, if gender is categorized as male = "1," female = "2," non-conforming = "3," non-binary = "4," and gender fluid = "5" in the left dataset, and as male = "M," female = "F," and other = "X" in the right dataset, then the gender values in the left dataset would be converted to match the categories in the right dataset. Specifically, male would be converted from "1" to "M," female from "2" to "F," and the values representing non-conforming, non-binary, and gender fluid would be converted to "X." This standardized form of sex/gender was then used as a blocking or matching variable in the data linkage algorithm.

Date fields were fragmented into separate day, month, and year fields by examining the date format (e.g., DD/MM/YYYY, MM/DD/YYYY, YYYY-MM-DD) and separating components accordingly. This decomposition allowed error tolerances to differ between date components. For example, birth years may be allowed to vary by at most one year, while exact matches may be required for birth month, and a larger margin of error allowed for the day of month (e.g., ± 3 days). Additionally, fragmented date fields enable matching the birth month in one dataset with the day of month in another to address data entry errors where these fields have been mistakenly swapped.

4.2.1 Missing Data Imputation

Missing sex/gender values were inferred from the primary given name (the first string in the given name field). In cases where multiple data elements were combined into single fields (i.e., compound fields), such as postal codes appended to address fields, regular expressions were used to extract data elements and reduce missingness in the respective fields.

4.3 Linkage Algorithm

Records were linked through an iterative process consisting of both deterministic and probabilistic steps. To improve the linkage rate, particularly for records with data entry errors in the fields used in previous passes, we varied the blocking and matching variables across iterations of the multi-step algorithm. Following each linkage step, matched record pairs were excluded from further consideration in subsequent iterations.

Initially, multiple deterministic linkage steps were conducted where agreement on matching variables was weighted equally. Linkage rates for deterministic passes were reported separately to allow data users to assess the reliability of matches. Specifically, record pairs that matched exactly on fields with high discriminative power are more likely to belong to the same individual, providing greater confidence in the linkage results.

For probabilistic linkage, we employed the Fellegi-Sunter model,[14] with several key extensions, to generate match weights for each candidate record pair. Blocking was employed to filter the Cartesian product of all possible candidate record pairs, improving computational feasibility by reducing memory requirements. The M and U conditional probabilities, which are necessary for estimating match weights, were calculated using the Expectation-Maximization (EM) unsupervised learning algorithm.[15] Identifier weights were adjusted based on observed frequencies, assigning lower weights to more frequent values (e.g., "John") and higher weights to less common values (e.g., "Barret"). Approximate string matching was performed using the Jaro-Winkler similarity metric to account for variations in character fields.[13] Strings with Jaro-Winkler similarity scores exceeding a selected threshold were considered a match. Total match weights, which are initially unbounded, were normalized to a [0,1] scale by converting them to posterior probabilities. This transformation to a standardized range was performed to enhance interpretation and facilitate comparisons across different passes of the linkage algorithm. Instead of the traditional two-threshold approach, a single acceptance threshold was used to classify candidate record pairs as matches or non-matches, and no clerical review of record pairs was performed.

Within each probabilistic pass, partial match weights were computed for each pair of matching variables, summed, and then normalized to produce a total match weight for each candidate record pair. Acceptance thresholds were determined by examining histograms of match weight distributions, stratified by true match status as defined by the ground truth, and manually selected to balance minimizing false positives and maximizing the linkage rate. Candidate record pairs exceeded the selected threshold were linked. In cases where a single record in the left dataset had multiple potential matches in the right dataset with similarity scores above the threshold, only the pair with the highest match weight was linked. If multiple potential matches in the right dataset were tied for the highest match weight, the pair with the most recent data acquisition date was selected.

4.4 Linkage Algorithm Evaluation

Linkage rates were stratified by sociodemographic characteristics (Table 2) to enable an examination of the algorithm's consistency across demographic groups and identify potential selection bias.

References

1. Bohensky MA, Jolley D, Sundararajan V, Evans S, Ibrahim J, Brand C. Development and validation of reporting guidelines for studies involving data linkage. *Australian and New Zealand journal of public health*. 2011;35(5):486-489. doi:10.1111/j.1753-6405.2011.00741.x

2. Gilbert R, Lafferty R, Hagger-Johnson G, et al. GUILD: Guidance for information about linking data sets. *Journal of Public Health*. 2017;40(1):191-198. doi:10.1093/pubmed/fdx037

3. Pratt NL, Mack CD, Meyer AM, et al. Data linkage in pharmacoepidemiology: A call for rigorous evaluation and reporting. *Pharmacoepidemiology and drug safety*. 2020;29(1):9-17. doi:10.1002/pds.4924

4. Elstad M, Ahmed S, Røislien J, Douiri A. Evaluation of the reported data linkage process and associated quality issues for linked routinely collected healthcare data in multimorbidity research: A systematic methodology review. *BMJ open*. 2023;13(5):e069212. doi:10.1136/bmjopen-2022-069212

5. Zhao Y, Jarrett M, McGail K, Hills B. A proposed approach for standardized reporting of data linkage processes and results. *International Journal of Population Data Science*. 2022;7(3). doi:10.23889/ijpds.v7i3.1962

6. Katz A, Enns J, Smith M, Burchill C, Turner K, Towns D. Population data centre profile: The manitoba centre for health policy. *International Journal of Population Data Science*. 2019;4(2). doi:10.23889/IJPDS.V4I2.1131

7. Blake HA, Sharples LD, Harron K, Meulen JH van der, Walker K. Linkage of multiple electronic health record datasets using a "spine linkage" approach compared with all "pairwise linkages." *International Journal of Epidemiology*. 2023;52(1):214-226. doi:10.1093/IJE/DYAC130

8. Government of Manitoba. Moving to manitoba. Published online 2023.

9. Manitoba Centre for Health Policy. Term: Registered manitoba population. Published online 2023.

10. Government of Manitoba. The personal health information act. Published online 2023.

11. Government of Manitoba. The freedom of information and protection of privacy act. Published online 2023.

12. Christen P, Christen P. Data matching systems. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Published online 2012:229-242. doi:10.1007/978-3-642-31164-2_10

13. Winkler WE. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. Published online 1990.

14. Fellegi IP, Sunter AB. A theory for record linkage. *Journal of the American Statistical Association*. 1969;64(328):1183-1210.

15. Winkler WE. *Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage*. US Bureau of the Census Washington, DC; 2000.

16. Enamorado T, Fifield B, Imai K. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*. 2019;113(2):353-371. doi:10.1017/S0003055418000783

17. Stoughton E, Monchka BA. *Linkrep: Data Linkage Quality Reports*.; 2024. https://github.com/CHIMB/linkrep

18. Chuchmach C, Monchka BA. *Datastan: Minimizing Linkage Response Time with User Friendly Standardizing Functions and Applications*.; 2024. https://github.com/CHIMB/datastan