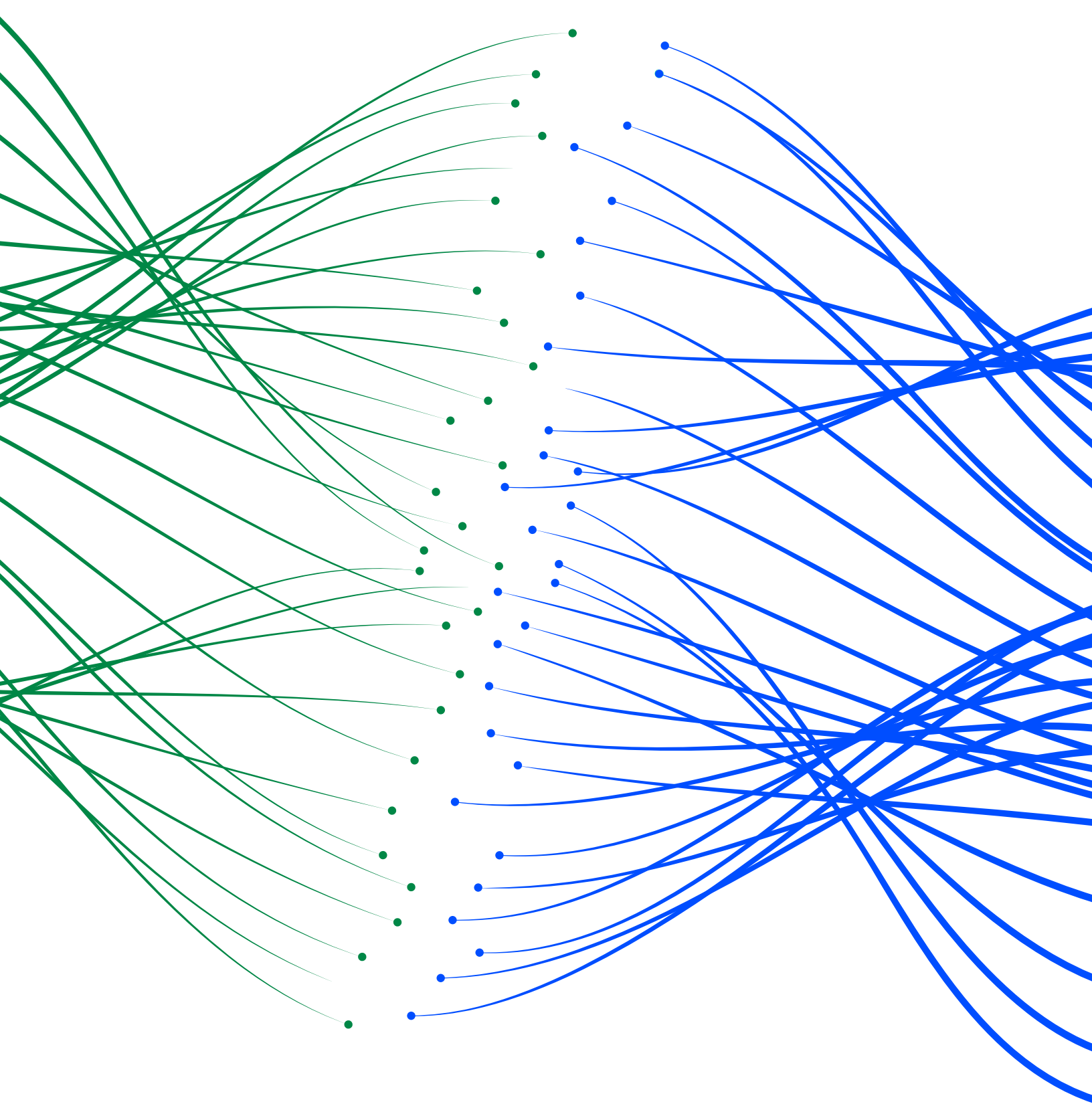


Data Linkage Sensitivity Analysis

Linkage of Synthetic Dataset A with Synthetic Dataset B

Prepared on Jun. 23, 2025

GEORGE & FAY YEE
Centre for Healthcare Innovation

University of Manitoba

Manitoba

This report was prepared by Elizabeth Stoughton and Barret Monchka at the George & Fay Yee Centre for Healthcare Innovation (CHI), in collaboration with the Manitoba Centre for Health Policy (MCHP) and the Provincial Information Management and Analytics (PIMA) department at Manitoba Health (MH). We gratefully acknowledge the contributions of Randy Walld of MCHP and Craig Kasper of MH for their roles in establishing the analytical and operational processes used to prepare the data for this report, and Karan Singh and Patrick Faucher of CHI Knowledge Translation for producing the graphics. Financial support was provided by CHI under the Canadian Institutes of Health Research (CIHR) Strategy for Patient Oriented Research (SPOR).

The software used to generate this report was developed by Elizabeth Stoughton and Cole Chuchmach, under the direction of Barret Monchka, and is available as an R package (https://github.com/CHIMB/linkrep) under an MIT license.

Permission is granted to adapt and reproduce the textual and graphical content of this report, in whole or in part, for non-commercial purposes related to generating data linkage quality reports, provided the original authors are appropriately attributed. If no textual or graphical content is reproduced in dynamically generated reports, then the copyright statement and suggested citation may be revised as necessary. Findings contained in this report may be reproduced provided the source is cited. All reasonable precautions have been taken by CHI to verify the accuracy of the information contained in this report, which is provided without warranty of any kind, either expressed or implied. The reader is responsible for the interpretation of the published material. In no event shall CHI, MCHP, MH, the University of Manitoba, the Government of Manitoba, or any data providers be liable for damages arising from the use of the published material or the report-generating software.

© 2025 George & Fay Yee Centre for Healthcare Innovation

Suggested Citation:

Elizabeth Stoughton, Cole Chuchmach, Barret Monchka. *Data Linkage Sensitivity Analysis: Linkage of Synthetic Dataset A with Synthetic Dataset B*. Winnipeg, Manitoba: George & Fay Yee Centre for Healthcare Innovation; June 2025.

Table of Contents

Definition of Terms

Positive predictive value (PPV) : Proportion of predicted positive matches that are truly positive

Negative predictive value (NPV) : Proportion of predicted negative matches that are truly negative

Sensitivity : Proportion of positive matches the algorithm correctly identified

Specificity : Proportion of negative matches the algorithm correctly identified

F1-score : A summary measure of the performance of the predicitve ability on the postitive class. Summarizes PPV and Sensitivity into a single number using a harmonic mean

Linkage rate : Proportion of source records that linked

Jaro-Winkler (JW) : Similarity metric for approximate string matching

False discovery rate (FDR) : Proportion of incorrect links among all detected links

False omission rate (FOR) : Proportion of missed correct links among all actual links

List of Abbreviations

CHI : George and Fay Yee Centre for Healthcare Innovation

MCHP : Manitoba Centre for Healthcare Policy

MH : Manitoba Health

PHIN : Personal Health Identification Number

Summary

This report details the methods used to link records from Synthetic Dataset A to those in Synthetic Dataset B and offers guidance on evaluating the quality of the linkage process. We used an iterative approach consisting of both deterministic and probabilistic matching techniques that achieved overall linkage rates ranging from 35.3% to 98.0% (see **Table 2**).

To discuss the findings of this report, the potential impact of linkage errors on a research study, or to request guidance on adjusting for linkage errors in a data analysis, you may schedule a consultation with CHI by visiting https://umanitoba.ca/centre-for-healthcare-innovation/ and selecting "Request a free consultation." For additional questions regarding this data linkage, please e-mail data.linkage@umanitoba.ca.

How to Read This Report

Integrating information from multiple data sources can generate rich and comprehensive datasets that address complex research questions. However, the uncertainty inherent in linking disparate data sources—particularly when there are no unique personal identifiers in common—may introduce bias, potentially reducing the representativeness of the linked data with respect to the target study population. This potential adverse impact on study validity has led to the development of reporting guidelines for studies using linked data, which have informed the elements included in this data linkage quality report.[1–5]

Researchers using this linked data should familiarize themselves with the record linkage methods employed, carefully assess the representativeness of the linked sample, and consider how linkage errors may impact the interpretation of study findings. The detailed information in the Background and Methods sections available online will aid in interpreting the results of the linkage process. When disseminating research based on linked data, we encourage adherence to reporting guidelines to support critical evaluations of findings and promote research excellence.[1–3]

The **Results** section includes figures and tables with descriptive statistics that summarize the performance of the linkage algorithm and the quality of the linked data. To assess the suitability of the data for research purposes, consider the following recommendations:

1. **Examine the representativeness of the linked sample:** Use the column percentages in **Table 1** to compare record characteristics with their distribution in the source data. Significant differences in these distributions may affect the generalizability of study findings.
2. **Assess potential biases in the linkage process:** Utilize the row percentages in **Table 2** to examine the proportion of unlinked records, stratified by sociodemographic and other characteristics. For example, a significantly lower linkage rate among females compared to males may indicate issues with the linkage algorithm or data quality that could introduce selection bias into a research study. Algorithmic biases can lead to an unrepresentative sample, as reflected in **Table 2**.
3. **Evaluate linkage algorithm accuracy:** Review **Table 6** and **Figure 1** to assess whether the linkage algorithm performed adequately in classifying candidate record pairs as matches or non-matches. Match classification, which involves a tradeoff between match certainty (i.e., precision or PPV) and match sensitivity (i.e., recall), should be assessed in the context of your research objectives.
4. **Critically review the multi-step linkage algorithm employed:** Analyze the parameters used in each step of the algorithm, including the matching technique (deterministic or probabilistic), acceptance threshold, and variables considered, to evaluate the confidence that linked records belong to the same individual. For example, exact matches on multiple fields generally reduce the chances of linkage error, whereas relying heavily on

approximate string matching may increase noise in the linked data. Similarly, higher acceptance thresholds and variables with strong discriminative power (e.g., last name or personal unique identifiers) tend to reduce linkage errors compared to lower classification cut-offs and variables with weaker discriminative ability (e.g., sex/gender or postal code).

1. Results

1.1 Linked Data Summary

Table 1: Characteristics of records in Synthetic Dataset A (N = 150) and those that linked to Synthetic Dataset B

	Algorithm 1	Algorithm 2	Algorithm 3	Source
	Linked (N = 53, 35.3%)	**Linked** (N = 97, 64.7%)	**Linked** (N = 147, 98.0%)	**Overall** (N = 150, 100.0%)
Sex				
Female	16 (30.2)	55 (56.7)	80 (54.4)	82 (54.7)
Male	37 (69.8)	42 (43.3)	67 (45.6)	68 (45.3)
Birth Year				
<1945	9 (17.0)	14 (14.4)	24 (16.3)	25 (16.7)
1945-1954	1 (1.9)	7 (7.2)	13 (8.8)	13 (8.7)
1955-1964	5 (9.4)	12 (12.4)	20 (13.6)	21 (14.0)
1965-1974	3 (5.7)	5 (5.2)	11 (7.5)	11 (7.3)
1975-1984	6 (11.3)	10 (10.3)	12 (8.2)	13 (8.7)
1985-1994	13 (24.5)	18 (18.6)	28 (19.0)	28 (18.7)
1995-2004	14 (26.4)	29 (29.9)	36 (24.5)	36 (24.0)
2005-2014	2 (3.8)	2 (2.1)	3 (2.0)	3 (2.0)
Geographic Region				

	Algorithm 1	Algorithm 2	Algorithm 3	Source
	Linked (N = 53, 35.3%)	**Linked** (N = 97, 64.7%)	**Linked** (N = 147, 98.0%)	**Overall** (N = 150, 100.0%)
Midwest	12 (22.6)	27 (27.8)	38 (25.9)	40 (26.7)
Northeast	13 (24.5)	23 (23.7)	39 (26.5)	40 (26.7)
South	14 (26.4)	22 (22.7)	40 (27.2)	40 (26.7)
West	14 (26.4)	25 (25.8)	30 (20.4)	30 (20.0)
Age				
<18	1 (1.9)	1 (1.0)	2 (1.4)	2 (1.3)
18-34	23 (43.4)	40 (41.2)	56 (38.1)	56 (37.3)
35-64	17 (32.1)	28 (28.9)	42 (28.6)	44 (29.3)
65-79	3 (5.7)	14 (14.4)	23 (15.6)	23 (15.3)
80+	9 (17.0)	14 (14.4)	24 (16.3)	25 (16.7)
Number of Given Names				
1	27 (50.9)	74 (76.3)	93 (63.3)	94 (62.7)
2	24 (45.3)	22 (22.7)	47 (32.0)	49 (32.7)
3+	2 (3.8)	1 (1.0)	7 (4.8)	7 (4.7)
Number of Surnames				
1	50 (94.3)	79 (81.4)	90 (61.2)	92 (61.3)
2	3 (5.7)	18 (18.6)	48 (32.7)	49 (32.7)
3+	0 (0.0)	0 (0.0)	9 (6.1)	9 (6.0)
Data Capture Year				

	Algorithm 1	Algorithm 2	Algorithm 3	Source
	Linked (N = 53, 35.3%)	**Linked** (N = 97, 64.7%)	**Linked** (N = 147, 98.0%)	**Overall** (N = 150, 100.0%)
2024	28 (52.8)	44 (45.4)	64 (43.5)	66 (44.0)
2025	25 (47.2)	53 (54.6)	83 (56.5)	84 (56.0)
Income Quintile				
1 (Lowest)	13 (24.5)	21 (21.6)	30 (20.4)	30 (20.0)
2	9 (17.0)	18 (18.6)	29 (19.7)	31 (20.7)
3	11 (20.8)	18 (18.6)	31 (21.1)	31 (20.7)
4	9 (17.0)	20 (20.6)	30 (20.4)	30 (20.0)
5 (Highest)	11 (20.8)	20 (20.6)	27 (18.4)	28 (18.7)
Residence Locality				
Rural	34 (64.2)	63 (64.9)	87 (59.2)	89 (59.3)
Urban	19 (35.8)	34 (35.1)	60 (40.8)	61 (40.7)

Data are presented as n (column %)

1.2 Linkage Rate Summary

Table 2: Stratified linkage rates for records in Synthetic Dataset A (N = 150) that linked to Synthetic Dataset B

	Algorithm 1	Algorithm 2	Algorithm 3	Source
	Linked (N = 53, 35.3%)	**Linked** (N = 97, 64.7%)	**Linked** (N = 147, 98.0%)	**Overall** (N = 150, 100.0%)
Sex				
Female	16 (19.5)	55 (67.1)	80 (97.6)	82
Male	37 (54.4)	42 (61.8)	67 (98.5)	68
Birth Year				
<1945	9 (36.0)	14 (56.0)	24 (96.0)	25
1945-1954	1 (7.7)	7 (53.8)	13 (100.0)	13
1955-1964	5 (23.8)	12 (57.1)	20 (95.2)	21
1965-1974	3 (27.3)	5 (45.5)	11 (100.0)	11
1975-1984	6 (46.2)	10 (76.9)	12 (92.3)	13
1985-1994	13 (46.4)	18 (64.3)	28 (100.0)	28
1995-2004	14 (38.9)	29 (80.6)	36 (100.0)	36
2005-2014	2 (66.7)	2 (66.7)	3 (100.0)	3
Geographic Region				
Midwest	12 (30.0)	27 (67.5)	38 (95.0)	40
Northeast	13 (32.5)	23 (57.5)	39 (97.5)	40

	Algorithm 1	Algorithm 2	Algorithm 3	Source
	Linked (N = 53, 35.3%)	**Linked** (N = 97, 64.7%)	**Linked** (N = 147, 98.0%)	**Overall** (N = 150, 100.0%)
South	14 (35.0)	22 (55.0)	40 (100.0)	40
West	14 (46.7)	25 (83.3)	30 (100.0)	30
Age				
<18	1 (50.0)	1 (50.0)	2 (100.0)	2
18-34	23 (41.1)	40 (71.4)	56 (100.0)	56
35-64	17 (38.6)	28 (63.6)	42 (95.5)	44
65-79	3 (13.0)	14 (60.9)	23 (100.0)	23
80+	9 (36.0)	14 (56.0)	24 (96.0)	25
Number of Given Names				
1	27 (28.7)	74 (78.7)	93 (98.9)	94
2	24 (49.0)	22 (44.9)	47 (95.9)	49
3+	2 (28.6)	1 (14.3)	7 (100.0)	7
Number of Surnames				
1	50 (54.3)	79 (85.9)	90 (97.8)	92
2	3 (6.1)	18 (36.7)	48 (98.0)	49
3+	0 (0.0)	0 (0.0)	9 (100.0)	9
Data Capture Year				
2024	28 (42.4)	44 (66.7)	64 (97.0)	66
2025	25 (29.8)	53 (63.1)	83 (98.8)	84

	Algorithm 1	Algorithm 2	Algorithm 3	Source
	Linked (N = 53, 35.3%)	**Linked** (N = 97, 64.7%)	**Linked** (N = 147, 98.0%)	**Overall** (N = 150, 100.0%)
Income Quintile				
1 (Lowest)	13 (43.3)	21 (70.0)	30 (100.0)	30
2	9 (29.0)	18 (58.1)	29 (93.5)	31
3	11 (35.5)	18 (58.1)	31 (100.0)	31
4	9 (30.0)	20 (66.7)	30 (100.0)	30
5 (Highest)	11 (39.3)	20 (71.4)	27 (96.4)	28
Residence Locality				
Rural	34 (38.2)	63 (70.8)	87 (97.8)	89
Urban	19 (31.1)	34 (55.7)	60 (98.4)	61

Data are presented as n (row %), where the percentage indicates the row-wise linkage rate

1.3 Summaries of Considered Linkage Algorithms

Table 3: Summary of the multi-step algorithm "Algorithm 1" for linking records in Synthetic Dataset A to those in Synthetic Dataset B

Step	Linkage Technique	Blocking Scheme	Matching Criteria	Acceptance Threshold	Linkage Rate (%)	Cumulative Linkage Rate (%)
1	D	Given Name, Surname	Address, Birth Year		2.7	2.7
2	D	Surname, Address	Zip Code		4.1	6.7
3	D	Given Name, Surname, Sex	Birth Month, Birth Year		12.1	18.0
4	P	Surname, Sex, Birth Month	Given Name (JW≥0.8)	Match Weight (-5.5)	21.1	35.3

D = deterministic linkage approach, P = probabilistic linkage approach, JW = Jaro-Winkler similarity score

Table 4: Summary of the multi-step algorithm "Algorithm 2" for linking records in Synthetic Dataset A to those in Synthetic Dataset B

Step	Linkage Technique	Blocking Scheme	Matching Criteria	Acceptance Threshold	Linkage Rate (%)	Cumulative Linkage Rate (%)
1	D	Given Name, Surname	Birth Month, Birth Year		14.7	14.7
2	D	Address, Zip Code	Given Name, Birth Month, Birth Year		10.2	23.3
3	D	Given Name, Surname	Birth Month, Birth Year		20.9	39.3

Step	Linkage Technique	Blocking Scheme	Matching Criteria	Acceptance Threshold	Linkage Rate (%)	Cumulative Linkage Rate (%)
4	D	Zip Code, Address	Given Name, Birth Month, Birth Year		11.0	46.0
5	P	Surname, Sex, Birth Month	Given Name (JW≥0.8)	Match Weight (-5.5)	34.6	64.7

D = deterministic linkage approach, P = probabilistic linkage approach, JW = Jaro-Winkler similarity score

Table 5: Summary of the multi-step algorithm "Algorithm 3" for linking records in Synthetic Dataset A to those in Synthetic Dataset B

Step	Linkage Technique	Blocking Scheme	Matching Criteria	Acceptance Threshold	Linkage Rate (%)	Cumulative Linkage Rate (%)
1	D	Given Name, Surname	Birth Month, Birth Year		14.7	14.7
2	D	Address, Zip Code	Birth Month, Birth Year		19.5	31.3
3	D	Given Name, Surname	Birth Month, Birth Year		23.3	47.3
4	D	Zip Code, Address	Birth Month, Birth Year		30.4	63.3
5	D	Person ID, Given Name - 1st Field Value (Left), Surname - 1st Field Value (Left)	Sex, Birth Year		25.5	72.7
6	D	Person ID, Given Name - 2nd Field Value (Left), Surname - 2nd Field Value (Left)	Sex, Birth Year		2.4	73.3

Step	Linkage Technique	Blocking Scheme	Matching Criteria	Acceptance Threshold	Linkage Rate (%)	Cumulative Linkage Rate (%)
7	D	Person ID, Given Name - 1st Field Value (Left), Surname - 2nd Field Value (Left)	Birth Day, Birth Month, Birth Year		30.0	81.3
8	D	Person ID, Given Name - 2nd Field Value (Left), Surname - 1st Field Value (Left)	Birth Day, Birth Month, Birth Year		28.6	86.7
9	P	Birth Day, Birth Month	Given Name (JW≥0.9), Surname (JW≥0.9)	Match Weight (0.15)	85.0	98.0

D = deterministic linkage approach, P = probabilistic linkage approach, JW = Jaro-Winkler similarity score

1.4 Classification Performance of Considered Linkage Algorithms

Table 6: Classification performance for linking records in Synthetic Dataset A to those in Synthetic Dataset B

Algorithm Name	Sensitivity	Specificity	PPV	NPV	F1 Score	Linkage Rate
Algorithm 1	73.9	75.0	96.2	25.0	83.6	35.3
Algorithm 2	77.4	83.3	99.0	15.2	86.9	64.7
Algorithm 3	97.3	95.8	99.3	85.2	98.3	98.0

PPV = Positive predictive value, NPV = Negative predictive value.

Classification performance was estimated among record pairs with non-missing values for Person ID and reported as percentages (%).

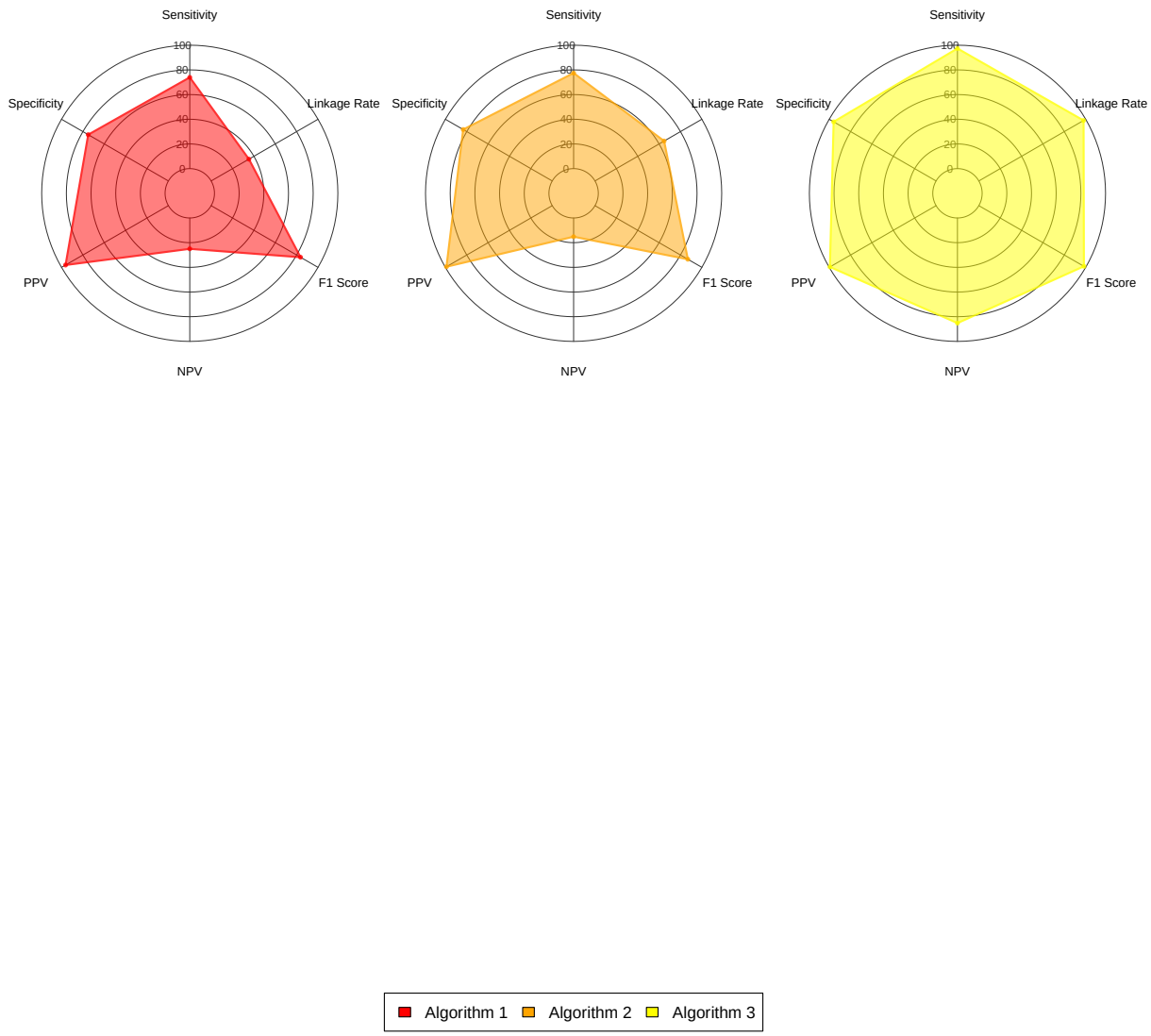


Figure 1: Radar charts showing classification performance for linking records in Synthetic Dataset A to those in Synthetic Dataset B. Classification performance was estimated among record pairs with non-missing values for Person ID and reported as percentages (%). PPV = Positive predictive value, NPV = Negative predictive value.

References

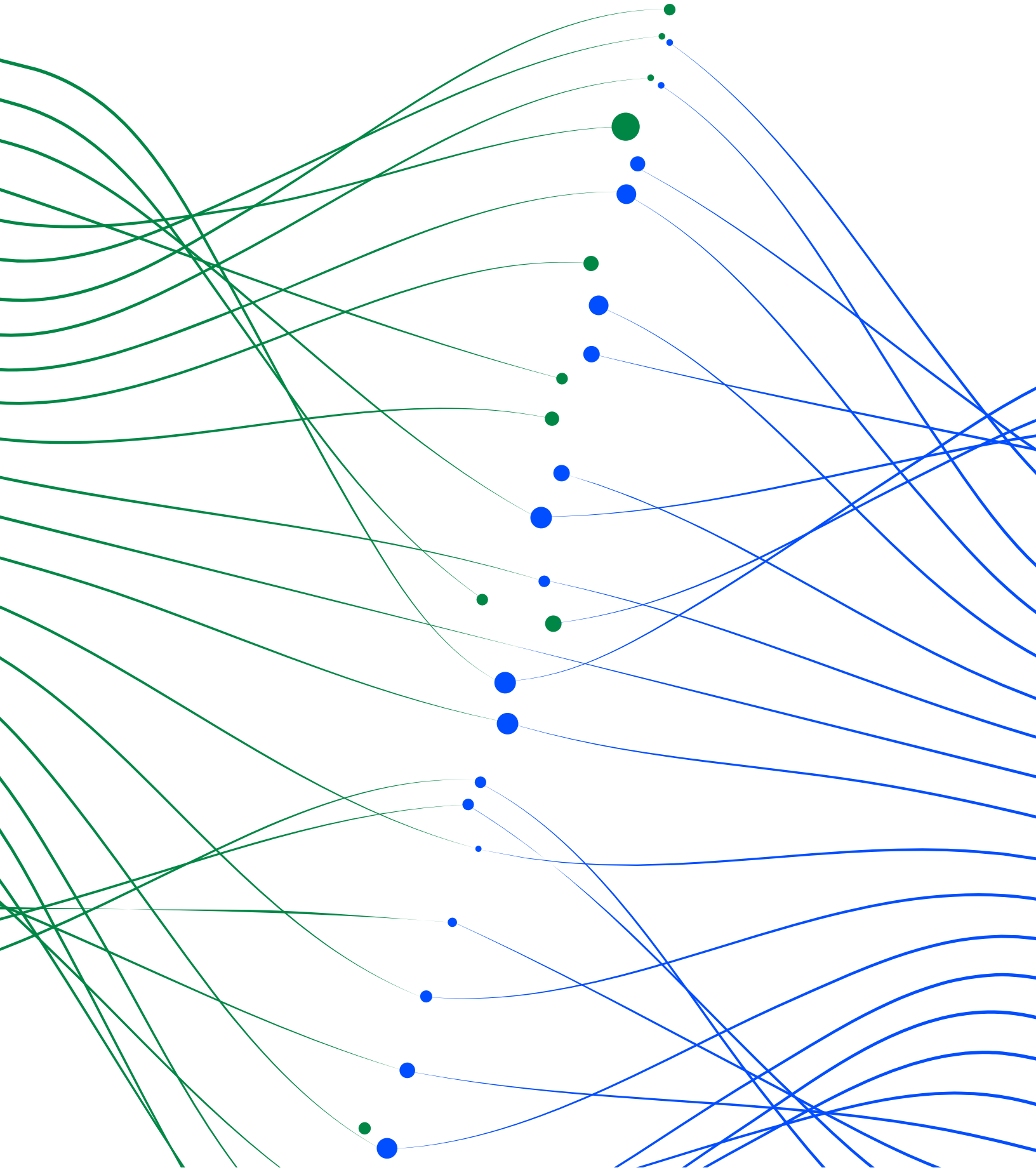
1. Bohensky MA, Jolley D, Sundararajan V, Evans S, Ibrahim J, Brand C. Development and validation of reporting guidelines for studies involving data linkage. *Australian and New Zealand journal of public health*. 2011;35(5):486-489. doi:10.1111/j.1753-6405.2011.00741.x

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



GEORGE & FAY YEE
Centre for Healthcare Innovation

University of Manitoba

Manitoba