International Journal of Current Research and Modern Education (IJCRME)
Impact Factor: 6.925, ISSN (Online): 2455 - 5428
(www.rdmodernresearch.com) Volume 7, Issue 2, 2022

PREDICTING PATIENT RECORD MATCHES USING DEMOGRAPHIC COMPARISON PATTERNS IN A GERMAN CANCER REGISTRY RESEARCH DATASET

Divya Vishwanath* & Paul Johnson Asamoah**
* Janardhan Rai Nagar Rajasthan Vidyapeeth, Udaipur, Rajasthan, India
** Power Life University College, Accra, Ghana, West Africa

Cite This Article: Divya Vishwanath & Paul Johnson Asamoah, "Predicting Patient Record Matches Using Demographic Comparison Patterns in a German Cancer Registry Research Dataset", International Journal of Current Research and Modern Education, Volume 7, Issue 2, Page Number 75-83, 2022.
Copy Right: © IJCRME, 2022 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract:

This analysis examines how demographic agreement relates to patient record match status, addressing the challenge of reliable health data linkage across settings. It uses 5,749,132 candidate pairs derived from German cancer registry records collected from 2005 to 2008. Reference match proportions are compared across agreement groups, with confidence intervals accounting for dependence between pairs that share a record. Exact agreement on both names was associated with a 66.540 percentage point higher match proportion, while complete birth date agreement and postal code agreement were associated with differences of 6.259 and 63.152 percentage points, respectively. These unadjusted comparisons support the tested associations within the selected candidate frame. However, most pairs with complete birth date agreement were reference nonmatches, showing that agreement can provide evidence without establishing identity. The findings illustrate the probabilistic principle that the value of agreement depends on its occurrence among matches and nonmatches. Health data managers can use this distinction to evaluate matching rules and review both false matches and missed matches before applying them in other settings.

Key Words: Candidate Pairs; Demographic Agreement; Health Data Linkage; Match Status; Probabilistic Linkage

1. Introduction:

Accurate identification of records belonging to the same person is essential for reliable health research. When records for one patient remain separate, the same person may be counted more than once. When records for different patients are combined, their information may be attributed to the wrong individual. These errors can distort disease estimates and analyses based on linked records (Harron et al., 2017; Moore et al., 2016).

Record linkage uses information shared across records to distinguish likely matches from nonmatches. The probative value of agreement depends on how often it occurs among records belonging to the same person and among records belonging to different people. Agreement on a common characteristic may occur by chance, while disagreement can arise from recording errors or changes in personal information. Probabilistic linkage formalizes these competing possibilities by treating comparison patterns as evidence rather than definitive proof of identity (Sayers et al., 2016).

Names, dates of birth, and postal codes provide different forms of evidence. First and family names may be similar across records for the same patient, but spelling differences and name changes can reduce agreement. Birth date components are generally stable, although different people can share a date and records can contain errors. Postal code agreement can provide information about location, but unrelated people may share a postal area and individuals may move. The expected relationship is therefore positive for each form of agreement, while the strength of that relationship depends on the frequency of coincidental agreement and the quality of the recorded data (Grannis et al., 2019; Karr et al., 2019).

The classification of a record pair as a match is also a measurement decision. Reference classifications can support comparisons of agreement patterns, but their value depends on how candidate pairs were selected and reviewed. Blocking reduces the number of pairs requiring comparison by selecting records that meet specified conditions. When those conditions use the same characteristics later examined as predictors, the observed associations apply to the selected candidate pairs. They cannot automatically be extended to every possible pair (Sariyar et al., 2012).

The closest research has examined linkage methods using the same German cancer registry data. Sariyar et al. (2012) evaluated approaches to missing comparison values using classification trees. Xu et al. (2019) examined whether accounting for dependence among comparison fields improves probabilistic linkage, including an application to the registry data. These studies establish that the dataset can inform methodological evaluation. They also show that conclusions depend on how missing information and relationships among fields are handled. The present analysis uses a narrower descriptive comparison: it examines how observed reference match proportions differ between agreement and disagreement groups for names, complete birth date, and postal code. It does not claim that these fields or their relevance are newly discovered.

This study aims to examine the associations between demographic agreement and reference match status among candidate pairs from a German cancer registry. The research question is: How do reference match proportions differ between candidate record pairs that agree and those that disagree on names, complete birth date, and postal code? The analysis addresses this question through three objectives: compare match proportions by joint name agreement, compare them by agreement across birth day, month, and year, and compare them by postal code agreement.

The dataset contains 5,749,132 candidate pairs derived from registry records collected from 2005 to 2008. It includes 20,931 reference matches and 5,728,201 reference nonmatches. The analysis compares observed match proportions within the available pairs; it does not estimate a prediction model or a causal effect. The reference classifications were informed by review, but the comparisons should not be treated as independent validation because information used in pair selection and review overlaps with the fields examined.

The analysis contributes a direct comparison of three agreement patterns against a common reference outcome within the available candidate frame. Its findings indicate that agreement on each selected field is associated with a higher reference match proportion, while agreement alone does not establish identity. This distinction can inform the evaluation of matching rules in other health data settings, subject to differences in candidate selection, data quality, and reference classification. Evidence from public health linkage also shows that matching methods can differ in their ability to identify true matches when records contain errors (Avoundjian et al., 2020).

2. Theory, Conceptual Framework and Hypotheses Development:

Record linkage depends on the ability to distinguish records belonging to the same person from records belonging to different people. Probabilistic linkage theory treats agreement between records as evidence whose value depends on how often that agreement occurs among both matching and nonmatching pairs. Agreement on a common characteristic may provide less evidence than agreement on a more distinctive characteristic. Sariyar, Borg, and Pommerening, 2011, show that linkage decisions must account for false matches as well as missed matches. Harron et al., 2017, explain that linkage errors can affect the validity of analyses based on linked records.

This analysis examines three observed dimensions of agreement: name, date of birth, and postal code. The outcome is the reference match status assigned to each candidate record pair through manual review. Name agreement is represented by similarity scores for the first components of first and family names. Date of birth agreement is represented by agreement on birth day, month, and year. Postal code agreement is represented by an indicator of whether the two postal codes agree. These are observed comparison measures rather than latent constructs. The expected relationships concern associations with reference match status, not the causal effect of changing a record.

Name Agreement and Record Match Status:

Name agreement captures the similarity of first and family names across two records. Records referring to the same person should generally have more similar names than records referring to different people. Differences in spelling, transcription, or name changes can nevertheless reduce agreement within true matches.
The value of a name comparison also depends on how frequently the name occurs. Shared names can produce agreement between different people, while partial similarity may reflect either a recording error or a coincidental resemblance. Name agreement therefore provides evidence of a match without establishing identity on its own.
Grannis et al., 2019, found that standardizing demographic fields, including family names, improved matching sensitivity in selected datasets. Karr et al., 2019, found that performance varied across linkage methods, particularly when inexact comparisons were used. These findings support a positive expectation while showing that the strength of the association depends on the comparison method and the data.

H1: Greater agreement on first and family names is associated with a higher probability that a candidate record pair is a reference match.

Date of Birth Agreement and Record Match Status:

Date of birth agreement reflects whether the day, month, and year recorded for two records correspond. Because these components describe the same person across records, agreement across them should increase the probability that a pair refers to one individual.

Unlike names, dates of birth do not normally change. Recording errors can still produce disagreement among true matches, and different people can share a birth date. The three components therefore provide complementary evidence, but none guarantees a match.

Weber et al., 2012, found that a linkage method combining name information with date of birth identified shared patients in their validation dataset. Karr et al., 2019, also showed that the use of date of birth in candidate selection affected linkage results. This analysis expects a positive association within the supplied candidate pairs, while recognizing that their original selection used some birth date information.

H2: Greater agreement across the day, month, and year of birth is associated with a higher probability that a candidate record pair is a reference match.

Postal Code Agreement and Record Match Status:

Postal code agreement captures whether two records contain the same postal code. It adds contextual evidence because records for the same person may contain the same residential location. It differs from name and date of birth agreement because location can change over time.

A shared postal code can also occur among unrelated people who live in the same area. Conversely, a person who moves may have different postal codes across records. These competing mechanisms make postal code agreement informative but potentially less stable as a marker of identity.

Harron et al., 2017, identify changes in characteristics such as postal code as a source of linkage error. Sariyar et al., 2012, show that missing comparison values also affect linkage procedures. Accordingly, the expected association applies to pairs for which postal code agreement is observed.

H3: Postal code agreement is associated with a higher probability that a candidate record pair is a reference match.

Conceptual Framework:

The framework specifies three observed predictors of reference match status: name agreement, date of birth agreement, and postal code agreement. Each proposed relationship represents a positive association within the candidate pairs supplied for analysis. The reference outcome reflects manual review and may contain classification error. The candidate pairs were selected through blocking rules that used some of the same characteristics assessed here, so the associations apply to the selected pairs and cannot be assumed to represent all possible record pairs.

The analysis can test all three hypotheses using the observed comparison fields and reference match status. Missing comparisons require an explicit treatment and should not automatically be interpreted as disagreement. Results may support, fail to support, or remain inconclusive for each hypothesis.

3. Data:

This section identifies the source and structure of the record pairs, documents the available sample, and defines the observed variables. It also explains how the data support the proposed analysis of demographic agreement and reference match status.

3.1 Data Sources, Population, and Sample Construction:

The analysis uses the Record Linkage Comparison Patterns dataset from the University of California, Irvine Machine Learning Repository. The underlying records originate from the epidemiological cancer registry of North Rhine Westphalia, Germany. The Institute for Medical Biostatistics, Epidemiology and Informatics at the University Medical Center Mainz created the comparison patterns and reference classifications. The repository lists the dataset under identifier 210 and provides it under a Creative Commons Attribution 4.0 International licence. The original download date and a separate version number are not recorded in the supplied files (Schmidtmann et al., 2011).

The source documentation describes a sample of 100,000 registry records collected from 2005 to 2008. The unit of analysis is a candidate pair of records. Each observation contains comparison measures and a reference classification indicating whether the pair relates to the same person. The data form a cross section of candidate pairs. Individual pair dates are unavailable, so annual changes cannot be estimated.

The target population comprises candidate record pairs generated from the sampled registry records. The accessible sampling frame consists of pairs selected by the data provider through six blocking procedures based on name, birth date, and sex agreement. The frame excludes pairs that did not satisfy any blocking rule. Manual review informed the reference classifications, although these classifications may contain error (Sariyar et al., 2012).

The ten supplied blocks contain 5,749,132 candidate pairs, including 20,931 reference matches and 5,728,201 reference nonmatches. The inspection identified 99,788 distinct record identifiers represented in the pairs. These identifiers refer to source records rather than distinct people.

Table 1: Sample construction

Stage	Record Identifiers	Candidate Pairs	Rule or Verified Status
Original source sample	100,000 records	Not reported	Registry records described in the source documentation
Accessible candidate frame	99,788 represented identifiers	5,749,132	Pairs selected by the source provider through blocking
Inspection for repeated ordered pairs	Not recalculated	5,749,132	No repeated ordered identifier pairs detected
Pairs with all proposed predictor fields observed	Not calculated	5,734,488	Potential complete case sample across the six predictor fields and match status

The complete case count is a coverage assessment rather than a final estimation sample. The supplied documentation does not establish when the original blocking rules were adopted or whether they changed during dataset preparation.

3.2 Data Integration, Cleaning, and Quality Assurance:

The supplied dataset comprises ten blocks with a common variable structure. The inspection read the blocks together as one collection of candidate pairs. The blocks were appended by rows, and no external dataset was merged. Two internal record identifiers identify the members of each pair. No duplicate ordered identifier pairs were detected. The inspection did not establish whether reversed identifier pairs occur.

The comparison fields use a question mark to denote missing values. The inspection treated this symbol as missing when calculating coverage and descriptive statistics. It did not replace missing values with zero, impute values, or remove observations from an estimation dataset. The source does not identify the reason for each missing value, so unavailable values cannot be reliably distinguished from structurally inapplicable values.

The first components of first and family name use similarity scores from zero to one. All inspected values fall within this range. Birth day, birth month, birth year, and postal code comparisons contain zero or one when observed. Zero indicates disagreement, and one indicates agreement. The source files contain comparison results rather than the underlying names, dates, or postal codes. The original entries therefore cannot be independently verified (Schmidtmann et al., 2011).

First name agreement has 1,007 missing values. Each birth date component has 795 missing values, and postal code agreement has 12,843. The second components of first and family names have 5,645,434 and 5,746,668 missing values, respectively. These second components are outside the proposed framework. The treatment of missing comparison values can affect linkage analysis, and its implications depend on the data and method used (Sariyar et al., 2012).

No outlier correction, external validation, or data revision procedure is documented in the supplied materials. Currency conversion and temporal alignment do not apply to the comparison measures.

3.3 Variable Measurement, Construction, and Validation:

The outcome is the supplied reference match status. The predictors are observed comparisons of names, date of birth, and postal code. Higher name scores indicate greater similarity. For birth date components and postal code, one indicates agreement and zero indicates disagreement. The framework does not require a latent construct or a composite index.

Table 2: Variable dictionary

Variable	Analytical Role	Definition and Source Identifier	Unit and Direction	Missing Value Status	Coverage and Limitation
Reference match status	Outcome	is_match; manual reference classification	One record pair; true denotes a reference match	No missing labels detected	Reference classification may contain review error
First name agreement	Name predictor	cmp_fname_c1; similarity of first name first components	Score from zero to one; higher means greater similarity	1,007 missing	Raw names are unavailable
Family name agreement	Name predictor	cmp_lname_c1; similarity of family name first components	Score from zero to one; higher means greater similarity	No missing values detected	Raw names are unavailable
Birth day agreement	Date of birth predictor	cmp_bd; agreement on birth day	Zero or one; one means agreement	795 missing	Does not establish agreement on the complete birth date
Birth month agreement	Date of birth predictor	cmp_bm; agreement on birth month	Zero or one; one means agreement	795 missing	Does not establish agreement on the complete birth date
Birth year agreement	Date of birth predictor	cmp_by; agreement on birth year	Zero or one; one means agreement	795 missing	Does not establish agreement on the complete birth date
Postal code agreement	Location predictor	cmp_plz; agreement on postal code	Zero or one; one means agreement	12,843 missing	Different people may share a postal code

All variables originate from the same repository dataset and have one observation per candidate pair. No transformations, weights, normalizations, interactions, or composite measures have been constructed.

Table 3 reports the supplied pairs before estimation exclusions. Standard deviations are sample standard deviations. For binary variables, the mean is the proportion of observed pairs with agreement.

Table 3: Descriptive statistics

Measure	Valid Pairs	Mean	Standard Deviation	Median	Minimum	Maximum
First name agreement	5,748,125	0.71290	0.38876	1.00000	0	1
Family name agreement	5,749,132	0.31563	0.33423	0.18182	0	1
Birth day agreement	5,748,337	0.22447	0.41723	0	0	1
Birth month agreement	5,748,337	0.48886	0.49988	0	0	1
Birth year agreement	5,748,337	0.22275	0.41609	0	0	1
Postal code agreement	5,736,289	0.00553	0.07415	0	0	1

These checks establish coverage and observed value ranges. They do not independently validate the reference labels or demonstrate the accuracy of a linkage model. The assessment of linkage quality depends on the reference data and evaluation design available (Harron et al., 2017).

3.4 Analytical Framework, Identification, and Reproducibility:

The intended analysis is associational. It examines the relationship between observed comparison fields and the probability that a candidate pair is classified as a reference match. The data support examination of the three proposed relationships involving name, date of birth, and postal code agreement. They do not support a causal interpretation.

The reference labels and predictor fields are not fully independent sources of information. The original review considered personal information and used clinical information in uncertain cases. This limits interpretation of the associations as independent confirmation of identity. The source blocking rules also selected pairs partly on the basis of characteristics available for analysis. Results would therefore apply to the selected candidate frame rather than all possible record pairs (Karr et al., 2019; Sariyar et al., 2012).

Candidate pairs can share an underlying record, so observations cannot automatically be treated as independent. The imbalance between reference matches and nonmatches also matters when interpreting model performance. False matches and missed matches represent different errors, and performance depends on the information and matching method used (Sariyar et al., 2011; Weber et al., 2012).

The repository DOI, source documentation, field identifiers, and coverage statistics provide a basis for reproducing the file inspection. Reproducible estimation requires documented code, a missing value rule, an estimator, and an inference procedure that addresses dependence between pairs. These procedures have not yet been documented.

4. Methods:

4.1 Research Design and Identification:

The analysis used an observational comparison of candidate record pairs to examine how demographic agreement relates to reference match status. The unit of analysis was one pair of registry records. For each agreement measure, the estimand was the difference in the proportion of reference matches between pairs with agreement and pairs in the corresponding comparison group.

The analysis relied on variation in the observed comparison fields within the candidate frame. It did not assign agreement conditions or identify a causal effect. The original blocking rules used personal characteristics to select candidate pairs, and reference review considered personal information. The resulting associations therefore describe the selected pairs and may partly reflect the procedures used to construct and classify them. They do not establish linkage accuracy across all possible record pairs or other health systems (Harron et al., 2017; Sariyar et al., 2012).

4.2 Comparison Measures and Estimands:

The variable dictionary in the Data section identifies the source fields and their coding. The analysis constructed three binary comparisons. For names, agreement required a similarity score of one for both the first and family name fields. The comparison group comprised pairs with a score below one on at least one of those fields. For date of birth, agreement required agreement on day, month, and year; the comparison group had disagreement on at least one component. For postal code, the analysis compared agreement with disagreement.
For each comparison, the estimated difference in match proportions was:

$$\widehat{\Delta}_k = \hat{p}_{k,1} - \hat{p}_{k,0}$$

In Equation 1, kidentifies the name, date of birth, or postal code comparison. The term $\hat{p}_{k,1}$ is the proportion of pairs classified as reference matches in the agreement group. The term $\hat{p}_{k,0}$is the corresponding proportion in the comparison group. A positive $\widehat{\Delta}_k$indicates a higher observed reference match proportion among pairs with agreement.

The three differences were estimated separately. No adjustment, interaction, composite index, or joint prediction model was fitted. The name contrast tests exact agreement against less than exact agreement; it does not test whether match probability increases at every level of the original similarity scores. Likewise, the date of birth contrast tests complete agreement against disagreement on at least one component. The contrasts were

constructed for this analysis and should be treated as exploratory rather than as documented prespecified comparisons.

4.3 Sample and Analytical Procedures:

The sample construction table in the Data section describes the original record sample and the blocked candidate frame. The supplied blocks contained 5,749,132 pairs. The analysis read the ten blocks together, treated the question mark as a missing value, and retained a pair for a comparison only when its reference label and all fields needed for that comparison were observed. Missing values were not replaced by disagreement and were not imputed.

The name comparison used 5,748,125 pairs. The date of birth comparison used 5,748,337 pairs. The postal code comparison used 5,736,289 pairs. The differences in sample size reflect comparison-specific missing values. No weights, outlier exclusions, or additional sample restrictions were applied.

The procedure counted reference matches and all observed pairs in each agreement group, calculated the group proportions, and subtracted the comparison group proportion from the agreement group proportion. Reference status was treated as the available outcome, not as an independently verified measure of identity. The source documentation reports manual review, including the use of clinical information in uncertain cases (Sariyar et al., 2012).

4.4 Inference, Validation, and Reproducibility:

A record can occur in more than one pair. The analysis therefore calculated standard errors allowing dependence between pairs that share a record identifier. This approach uses a dyadic variance calculation based on contributions from both members of each pair. It assumes that pairs without a shared identifier are independent. Dependence beyond shared records is not addressed by this calculation. Methods for dyadic data demonstrate why treating all pairs as independent can understate uncertainty (Aronow et al., 2015).

For each comparison, let u_i denote the contribution of pair i to the estimated difference. It equals the pair's reference match indicator minus its group match proportion, divided by the size of its group, with a negative sign for the comparison group. For each record identifier r, let S_r be the sum of u_i across pairs containing that identifier. The calculated variance was:

$$\hat{V}_k = \sum_r S_r^2 - \sum_i u_i^2$$

Equation 2 sums the record-level contributions and subtracts one copy of each pair's squared contribution because each pair contributes to two record sums. The standard error was the square root of the calculated variance, with negative calculated variances set to zero. The reported 95 percent intervals used the estimated difference plus or minus 1.96 standard errors. Their interpretation depends on the stated dependence assumption and the adequacy of the normal approximation.

The inspection checked observed value ranges and missingness and found no repeated ordered identifier pairs. It did not establish whether reversed identifier pairs occur. No alternative inference procedure, independent reference validation, or sensitivity analysis was performed. Accordingly, the confidence intervals do not address uncertainty from blocking, reference classification, or possible dependence between pairs without a shared identifier.

The calculations used Python 3.12.14, pandas 2.2.3, and NumPy 2.3.5. The source dataset and field definitions are available through the UCI Machine Learning Repository under dataset identifier 210 (Schmidtmann et al., 2011). The analysis script used for the reported comparisons exists in the working environment, but a persistent replication package has not been deposited. No preregistration was documented.

5. Findings:

The analysis examines whether agreement on names, date of birth, and postal code is associated with reference match status. The findings compare observed match proportions across candidate pairs and assess the three proposed hypotheses. The estimates describe associations within the selected candidate frame.

5.1 Sample Characteristics:

The ten data blocks contain 5,749,132 candidate record pairs from the German cancer registry dataset. Of these, 20,931 pairs, or 0.364 percent, are classified as reference matches. The remaining 5,728,201 pairs are classified as reference nonmatches. The underlying registry records cover 2005 to 2008.

The analytical sample varies by comparison because some fields are missing. The name analysis includes 5,748,125 pairs with both name comparisons observed. The date of birth analysis includes 5,748,337 pairs with all three birth date components observed. The postal code analysis includes 5,736,289 pairs with observed postal code comparisons.

The outcome is the supplied reference classification. Name agreement compares pairs with exact agreement on both name fields against pairs with less than exact agreement on at least one field. Date of birth agreement compares pairs with agreement on all three components against pairs with disagreement on at least one component. Postal code agreement compares pairs with equal postal codes against pairs with different postal codes.

5.2 Associations Between Agreement and Reference Match Status:

Table 1 presents the reference match proportion for each comparison. Confidence intervals concern the difference in match proportions between agreement groups. The interval calculation accounts for dependence between pairs that share a record, under the assumption that pairs without a shared record are independent. The comparisons are unadjusted and were constructed for this analysis.

Table 1: Demographic agreement and reference match status

Comparison	Agreement Group: Matches And Pairs	Match Proportion	Comparison Group: Matches and Pairs	Match Proportion	Difference in Percentage Points	95 Percent Confidence Interval
Exact agreement on both names	20,525 of 30,843	66.547%	397 of 5,717,282	0.0069%	66.540	65.051 to 68.028
Agreement on all birth date components	20,766 of 331,637	6.262%	159 of 5,416,700	0.0029%	6.259	6.140 to 6.377
Postal code agreement	20,033 of 31,714	63.168%	869 of 5,704,575	0.0152%	63.152	62.339 to 63.966

Name Agreement and Reference Match Status:

The reference match proportion was 66.547 percent among pairs with exact agreement on both names, compared with 0.0069 percent among pairs with less than exact agreement on at least one name. The difference was 66.540 percentage points, with a confidence interval from 65.051 to 68.028 percentage points.

This result supports H1 for the tested contrast between exact agreement on both names and less than exact agreement. It does not establish that every increase in name similarity is associated with a higher match probability, because the analysis compares two groups rather than the full range of similarity scores. The result is consistent with evidence that name information contributes to patient matching, although performance depends on the comparison method and the quality of recorded information, as shown by Karr et al., 2019, and Weber et al., 2012.

Date of Birth Agreement and Reference Match Status:

The reference match proportion was 6.262 percent among pairs agreeing on birth day, month, and year. It was 0.0029 percent among pairs with disagreement on at least one component. The difference was 6.259 percentage points, with a confidence interval from 6.140 to 6.377 percentage points.

The positive contrast supports H2 for complete birth date agreement relative to disagreement on at least one component. Nevertheless, 93.738 percent of pairs with complete birth date agreement were reference nonmatches. Complete agreement is therefore informative within this candidate frame but does not establish that two records belong to the same person. The original blocking procedures used birth date information, which affects the composition of pairs available for comparison, as described by Sariyar et al., 2012.

Postal Code Agreement and Reference Match Status:

The reference match proportion was 63.168 percent among pairs with postal code agreement, compared with 0.0152 percent among pairs with postal code disagreement. The difference was 63.152 percentage points, with a confidence interval from 62.339 to 63.966 percentage points.

This positive contrast supports H3 within the observed candidate pairs. Postal code agreement was not sufficient to establish a match: 11,681 pairs with postal code agreement were classified as reference nonmatches. Conversely, 869 pairs with postal code disagreement were classified as reference matches. These exceptions are consistent with the possibility that different people share a postal area and that a person's recorded location changes, a limitation discussed by Harron et al., 2017.

5.3 Interpretation of the Hypothesis Tests:

All three observed contrasts have positive estimated differences, and their confidence intervals exclude zero under the stated dependence assumption. The findings support H1 and H2 for the specific exact agreement contrasts tested and support H3 for observed postal code agreement. These are unadjusted comparisons. They do not establish the separate contribution of each field after accounting for the other fields.

The magnitude of each contrast is specific to the blocked candidate frame and the supplied reference classifications. The original selection rules used some of the same characteristics examined here, and manual review drew on personal information. Consequently, these estimates cannot be interpreted as independently validated performance in another population. Reference labels may also contain error. Sariyar et al., 2011, distinguish false matches from missed matches, while Harron et al., 2017, emphasize the importance of the reference data used to assess linkage quality.

The analysis identifies substantial associations between demographic agreement and reference match status in the available pairs. It does not establish causal effects, validate privacy protection, or demonstrate performance across national health systems.

6. Discussion:

Agreement between demographic fields provides evidence of a shared identity, but its value depends on the candidate pairs being compared. In Table 1, exact agreement on both names and agreement on postal code were associated with substantially higher reference match proportions than their respective comparison groups. Complete birth date agreement also showed a positive association, although most pairs agreeing on all birth date components were reference nonmatches. These contrasts illustrate how agreement can be informative without being sufficient for identification. Their contribution lies in placing the three comparisons on a common reference outcome within the available candidate frame. The data do not establish that these patterns are previously undocumented. The interpretation follows the probabilistic principle that the evidential value of agreement depends on how often it occurs among matching and nonmatching pairs (Sayers et al., 2016).

The contrast for complete birth date agreement has a practical implication. Although agreement across day, month, and year was associated with a higher reference match proportion, approximately 94 percent of pairs with complete agreement were reference nonmatches. A linkage decision based on that agreement alone would therefore include many pairs assigned to different people by the reference classification. The result should be interpreted within the blocked candidate frame rather than as a universal property of birth dates. It is also an unadjusted comparison: the analysis does not establish the contribution of birth date agreement after accounting for names or postal code. Research on conditional dependence shows why information shared across comparison fields can affect the performance of linkage models (Daggy et al., 2014).

The stronger observed contrasts for joint name agreement and postal code agreement do not establish that either field dominates a causal pathway. The comparisons use different agreement definitions and slightly different samples, and the analysis did not estimate mediation or decompose a total effect. Their value is instead diagnostic: they show that agreement patterns and match proportions differ across fields. A theoretical explanation is that fields vary in their frequency of coincidental agreement and their susceptibility to recording differences. That explanation remains an interpretation rather than a mechanism identified by these estimates. Models that account for dependence among fields can produce different linkage results from models that treat the fields as independent (Xu et al., 2019).

The candidate selection process is central to interpreting the results. Blocking rules used name, birth date, and sex information before the supplied pairs were formed. The reference review also considered personal information. The observed associations therefore describe pairs selected and classified through a process related to the variables being examined. They cannot be transferred directly to all possible registry pairs or treated as independent validation of a linkage system. The international consequence is that a high match proportion in one selected set does not by itself establish accuracy in a different health system. In Australian cohort data, Moore et al. (2016) showed that linkage sensitivity and selection affected estimated outcomes, demonstrating why linkage quality must be considered alongside the analysis that uses linked records.

For practice, the findings support evaluating both agreements and disagreements before adopting a rule that links records. The postal code results show why: some pairs with equal postal codes were reference nonmatches, while some pairs with different postal codes were reference matches. In United States public health data, Avoundjian et al. (2020) found that linkage methods differed in their ability to recover matches when data quality varied. Ong et al. (2020) reported improved performance from a combined linkage approach in a separate health surveillance setting. Those comparisons suggest useful directions for evaluation, but they do not establish that either approach would perform better on the present data. Future analysis can test that question using an independently assessed reference sample, examine agreement fields jointly, and report both false matches and missed matches.

7. Conclusion and Implications:

Reliable health information depends on recognizing when separate records refer to the same person. This analysis shows that different forms of demographic agreement are associated with reference match status, while agreement alone does not establish identity. Its contribution is to compare these associations against a common reference outcome in a large set of candidate pairs. The findings reinforce a central principle of probabilistic linkage: the value of agreement depends on its occurrence among both matching and nonmatching records. They also show why linkage decisions require attention to false matches as well as missed matches. Health data managers can use this insight to assess proposed matching rules against reviewed pairs and examine cases where comparison fields disagree. Registry operators can document candidate selection and review procedures so that users understand the population to which linkage results apply. These practices can support more reliable disease measurement and research using health records. Their relevance extends beyond the registry examined here, although the observed proportions should not be assumed to apply in other settings.

The findings concern candidate pairs selected through blocking rules in one German cancer registry. Reference classifications were informed by information related to the comparisons, and the analysis assessed each association separately. Future research can evaluate the fields jointly using independently reviewed pairs, compare alternative selection rules, and test performance in registries from other countries. Such work can

assess whether the observed patterns transfer across populations and data systems and whether matching decisions remain reliable when recording quality differs.

References:

1. Aronow, P. M., Samii, C., & Assenova, V. A. (2015). Cluster-robust variance estimation for dyadic data. Political Analysis, 23(4), 564-577. https://doi.org/10.1093/pan/mpv018
2. Avoundjian, T., Dombrowski, J. C., Golden, M. R., Hughes, J. P., Guthrie, B. L., Baseman, J., &Sadinle, M. (2020). Comparing methods for record linkage for public health action: Matching algorithm validation study. JMIR Public Health and Surveillance, 6(2), e15917. https://doi.org/10.2196/15917
3. Daggy, J., Xu, H., Hui, S., & Grannis, S. (2014). Evaluating latent class models with conditional dependence in record linkage. Statistics in Medicine, 33(24), 4250-4265. https://doi.org/10.1002/sim.6230
4. Grannis, S. J., Xu, H., Vest, J. R., Kasthurirathne, S., Bo, N., Moscovitch, B., Torkzadeh, R., & Rising, J. (2019). Evaluating the effect of data standardization and validation on patient matching accuracy. Journal of the American Medical Informatics Association, 26(5), 447-456. https://doi.org/10.1093/jamia/ocy191
5. Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology, 46(5), 1699-1710. https://doi.org/10.1093/ije/dyx177
6. Karr, A. F., Taylor, M. T., West, S. L., Setoguchi, S., Kou, T. D., Gerhard, T., & Horton, D. B. (2019). Comparing record linkage software programs and algorithms using real-world data. PLOS ONE, 14(9), e0221459. https://doi.org/10.1371/journal.pone.0221459
7. Moore, C. L., Gidding, H. F., Law, M. G., Amin, J., & Han, A. (2016). Poor record linkage sensitivity biased outcomes in a linked cohort analysis. Journal of Clinical Epidemiology, 75, 70-77. https://doi.org/10.1016/j.jclinepi.2016.01.023
8. Ong, T. C., Duca, L. M., Kahn, M. G., & Crume, T. L. (2020). A hybrid approach to record linkage using a combination of deterministic and probabilistic methodology. Journal of the American Medical Informatics Association, 27(4), 505-513. https://doi.org/10.1093/jamia/ocz232
9. Sariyar, M., Borg, A., & Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. Journal of Biomedical Informatics, 44(4), 648-654. https://doi.org/10.1016/j.jbi.2011.02.008
10. Sariyar, M., Borg, A., & Pommerening, K. (2012). Missing values in deduplication of electronic patient data. Journal of the American Medical Informatics Association, 19(e1), e76-e82. https://doi.org/10.1136/amiajnl-2011-000461
11. Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. International Journal of Epidemiology, 45(3), 954-964. https://doi.org/10.1093/ije/dyv322
12. Schmidtmann, I., Hammer, G., Sariyar, M., & Gerhold-Ay, A. (2011). Record linkage comparison patterns [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C51K6B
13. Weber, S. C., Lowe, H., Das, A., & Ferris, T. (2012). A simple heuristic for blindfolded record linkage. Journal of the American Medical Informatics Association, 19(e1), e157-e161. https://doi.org/10.1136/amiajnl-2011-000329
14. Xu, H., Li, X., Shen, C., Hui, S. L., & Grannis, S. (2019). Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? The Annals of Applied Statistics, 13(3), 1753-1790. https://doi.org/10.1214/19-AOAS1256