Investigating the Problem of Patient Record Duplications in Health Information Systems – The Case of UL Hospitals

Abdulhussain E. Mahdi[1], Arash Joorabchi[1], and Brian McKeon[2]

[1] TAKO, Department of Electronic & Computer Engineering, University of Limerick; [2] eHealth Division, UL Hospitals Group

This paper reports on the findings of the exploratory phase of a joint research project between the TAKO (Text Analytics & Knowledge Organisation) Research Group - UL and the eHealth Division - ULH Group to investigate the extent of record duplication problem in the ULH's patients database and the potential solutions.

Duplicate Medical Records

A duplicate medical record occurs when a single patient is associated with more than one medical record.

Causes:

- The use of multiple information systems for clinical and administrative services
- Small errors and inconsistencies introduced mainly during the registration process.

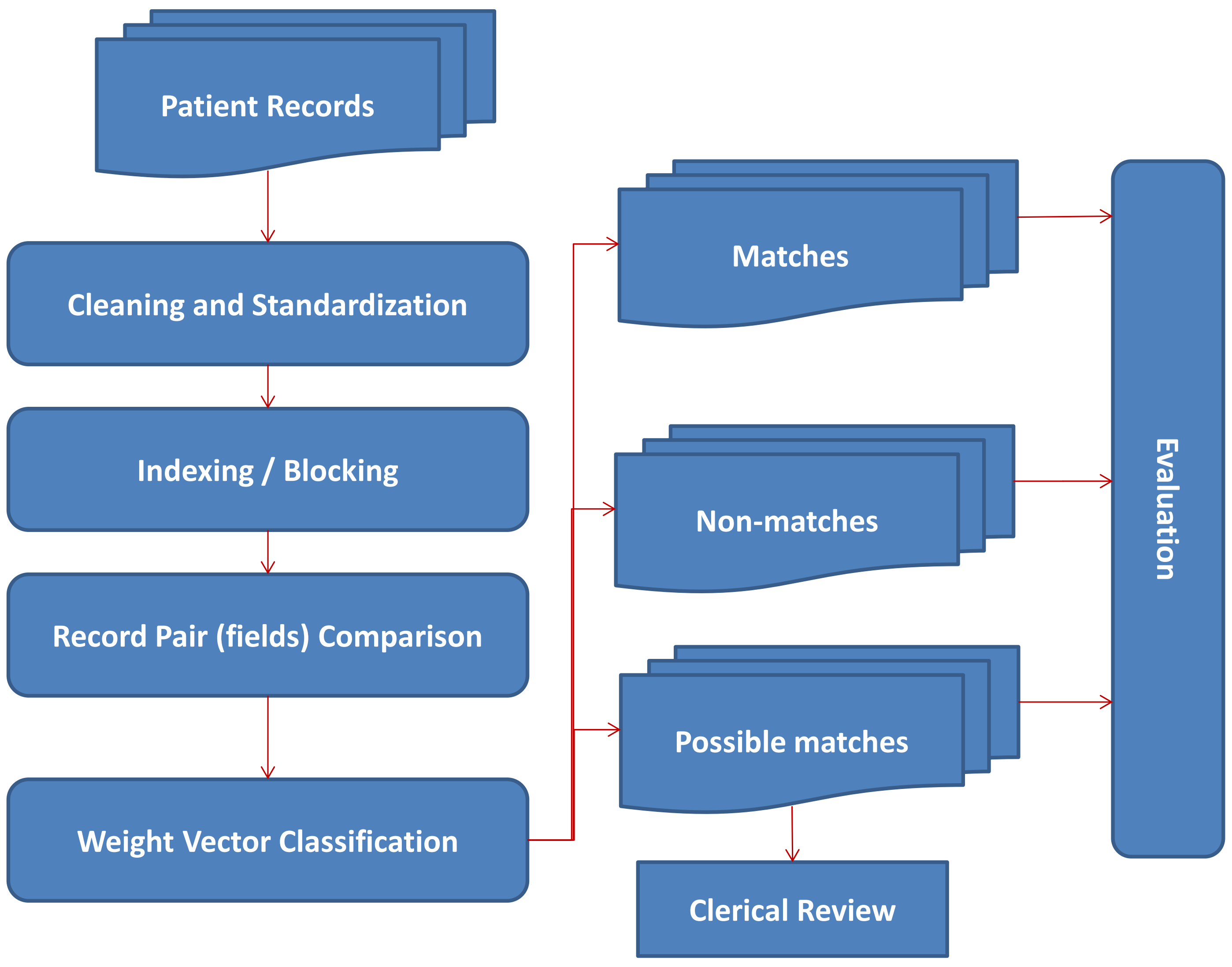
Symptoms:

- Partial records that only capture a portion of a patients medical history
- Treating patients based on incomplete medical history
- communications issues between healthcare providers and patients

Costs:

- Reported average duplication rate in American hospitals is 8% - 12%.
- The annual operational cost of a duplicate pair can be ~ €50.00 per pair.
- Unnecessary repeated tests and subsequent delays in starting the treatment result in an average of ~ € 1,100 extra cost per record.

Automatic Matching & Deduplication


Patient Records

Cleaning and Standardization

Indexing / Blocking

Record Pair (fields) Comparison

Weight Vector Classification

Matches

Non-matches

Possible matches

Clerical Review

Evaluation

1. **Cleaning & Standardization:**
   - Removal of unwanted characters and words
   - Expand abbreviations and correct misspellings
   - Attribute segmentation, e.g. breaking addresses to street, town, county
   - Verify the correctness of attribute values via external databases
2. **Indexing**: For a dataset containing *n* records, *n(n − 1)* comparisons have to be conducted. Hence a dataset containing 100,000 records would require 9,999,900,000 record pair comparisons. Indexing/blocking addresses this issue by splitting the records into smaller blocks according to defined criteria.
3. **Record Pair Comparison:** The various attributes of candidate records generated in above indexing step are compared to determine their similarities. For attributes that contain string values, e.g. names and addresses, a number of approximate string comparison functions are applied. Specific comparison functions for dates, ages, times, locations and numerical values are used for attributes that contain such data.
4. **Record Pair Classification:** A two-class classification process is applied. This has been achieved using a traditional probabilistic method, known as Fellegi-Sunter.
5. **Evaluation:** The accuracy of the classification of record pairs into matching and non-matching is evaluated using standard measures of Precision (Pr), Recall (Re), and their harmonic mean, F1:

$$\mathrm{Pr} = \frac{\text{Number of correctly detected duplicates}}{\text{Total detected}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{Re} = \frac{\text{Number of correctly detected duplicates}}{\text{Total true duplicates}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad \mathrm{F1} = \frac{2\mathrm{Pr}\times\mathrm{Re}}{\mathrm{Pre}+\mathrm{Re}}$$

ULH Master Patient INDEX

The database holding the patient records from the six hospitals in the ULH network contains over 1 million records. This Master Patient Index (MPI) was created by merging the patient records from 6 hospitals in August of 2015. We tracked an average of 177 new records per day over a 90 days period.

Patient Data Model

Patients: 1,067,365

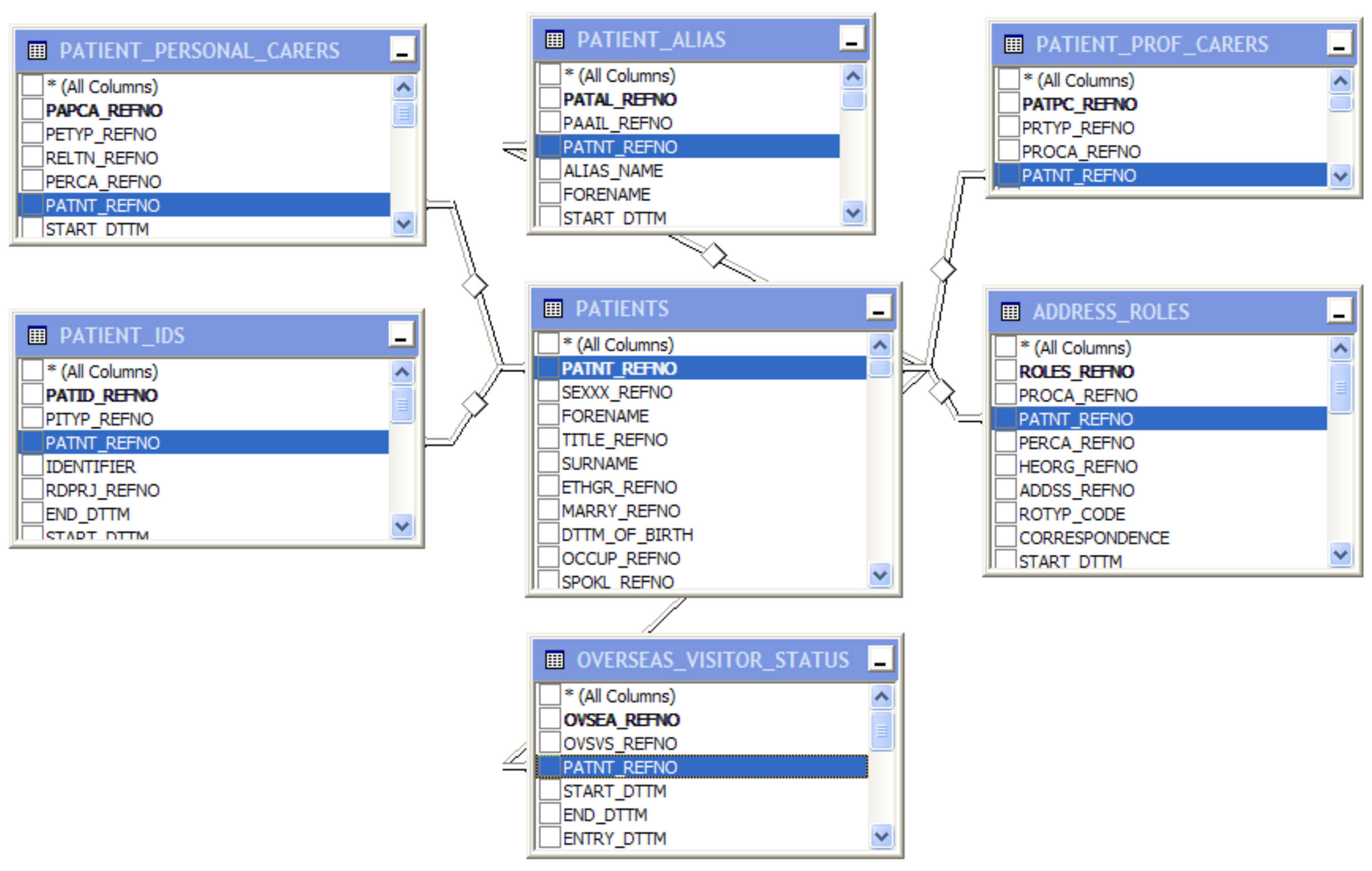
Female: 552,492

Male: 513,294

Unique Forenames:44,310

Unique Surnames: 95,497

Unique DOBs: 42,102



Address Data Model

Phones: 360,491

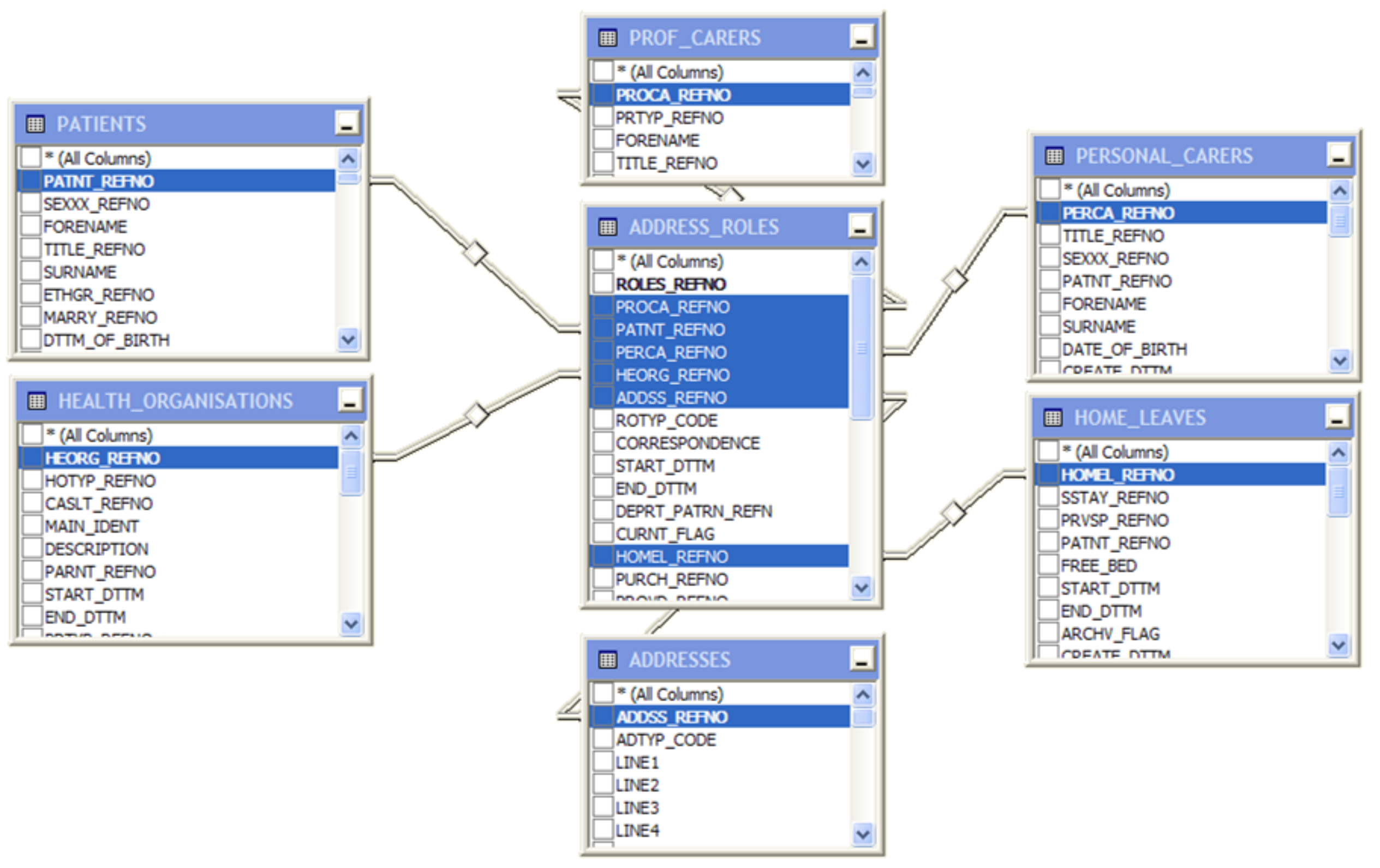
Mobiles: 404,981

Post Line 1: 1,065,303

Post Line 2: 1,061,267

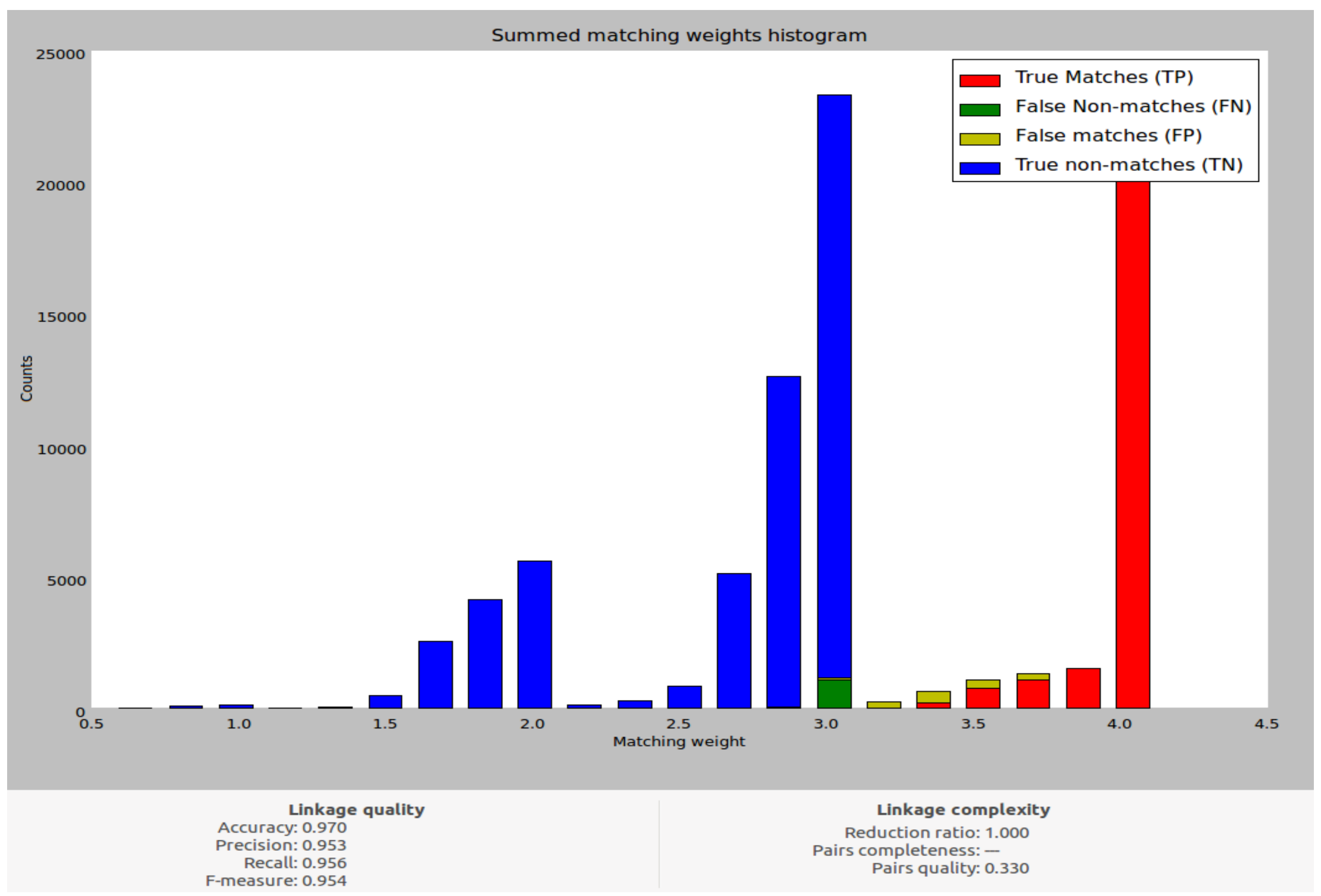
Post Line 3: 988,555

Post Line 4: 251,174



Experimental Deduplication Results

The accuracy of the deduplication system was first measures using a test collection of 50,000 records which are manually deduplicated by ULH hospital admins.



Linkage quality

Accuracy: 0.970

Precision: 0.953

Recall: 0.956

F-measure: 0.954

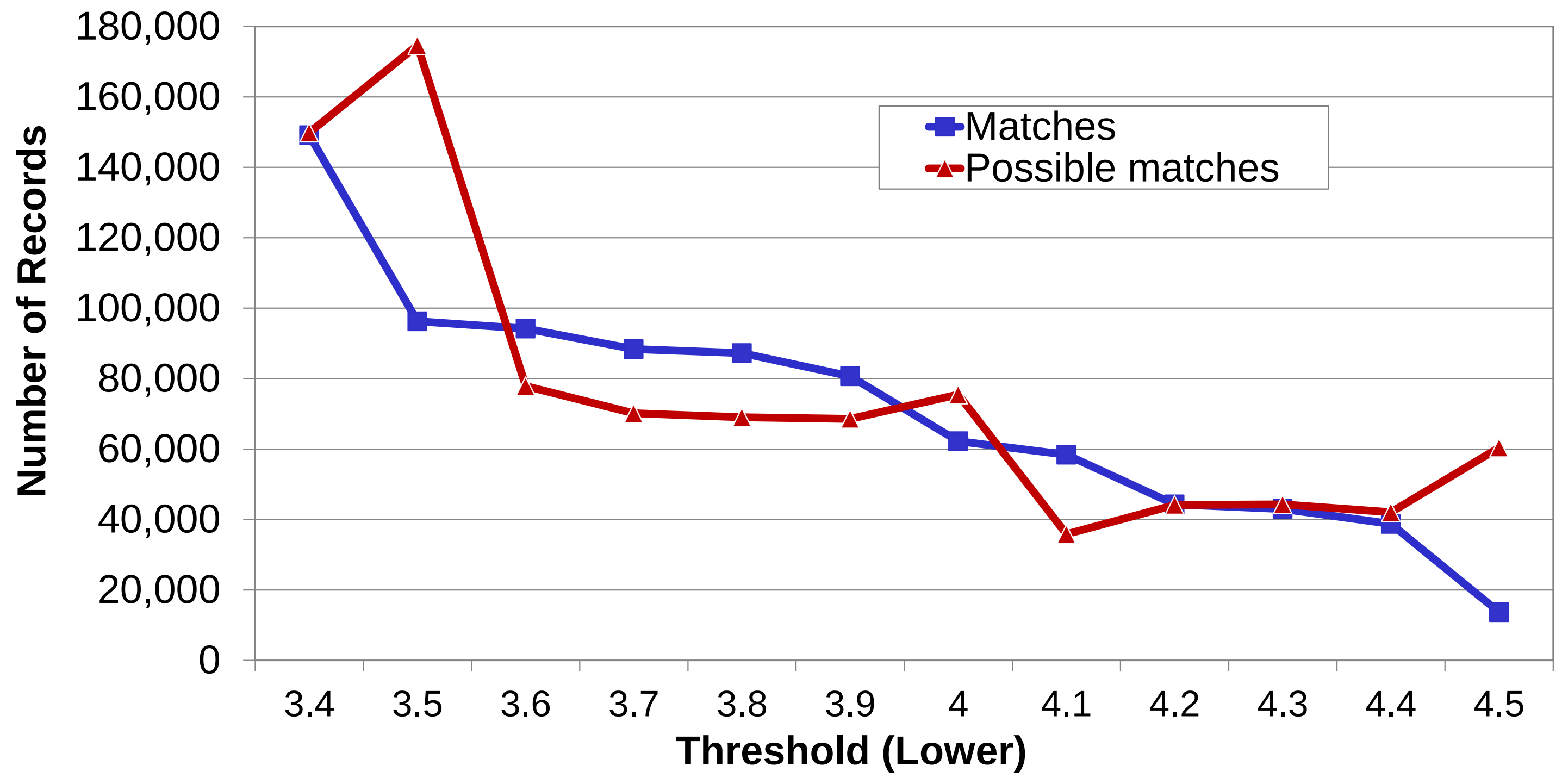
Linkage complexity

Reduction ratio: 1.000

Pairs completeness: —

Pairs quality: 0.330

After qualifying its accuracy, the system was applied to ULH's full database to gauge the level of duplication in the ULH-DB. A duplication rate of 4%-12% depending on the level of confidence (i.e., set classification thresholds) was detected.



Conclusions: we believe the findings of Phase 1 of this investigation have provided an insight into the nature and extent of the duplications in ULH-DB, and that our developed deduplication prototype can contribute greatly to improving the quality and integrity of the ULH patients' data. This can be achieved by further enhancement and integration of this prototype into ULH existing patients' information system.