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 patient's 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

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:

   \[
   \text{Pr} = \frac{\text{Number of correctly detected matches}}{\text{Total detected}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \\
   \text{Re} = \frac{\text{Number of correctly detected matches}}{\text{Total true matches}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \\
   \text{F1} = \frac{2\times \text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}}
   \]

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

After qualifying its accuracy, the system was applied to ULH’s full database to gauge the level of duplication in the ULH-DN. 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-DN, 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.