PCD leakage detection and redaction
PCD Leakage Detection

- As part of interchange validation SUS+ looks through records to detect PCD leakages.
- PCD leakage occurs when:
  - Value of a PCD field (e.g. NHS number) was found in a non-PCD field (e.g. provider reference number);
  - Value of a PCD field was found in another PCD field (e.g. patient pathway id);
- SUS+ currently provide reports results from the interchange validation.
Next steps

- Evaluated methods to identify false positives
  - The likelihood that a false positive has occurred
- Elaborated where in the processing to apply redaction
- Sharing the results before starting to implement the solution
True positive vs. false positive leakage

- **False positive**
  - 5 records out of 1000 contains date of birth in critical care local identifier:

```xml
<ns:PersonBirthDate>1959-02-09</ns:PersonBirthDate>
...  
<ns:CriticalCareLocalIdentifier>1302095980</ns:CriticalCareLocalIdentifier>
```

- **True positive**
  - 500 records out of 1000 contain date of birth in critical care local identifier
  - 30 records out of 1000 contain NHS number in critical care local identifier

```xml
<ns:NHSNumber>1234567890</ns:NHSNumber>
...  
<ns:CriticalCareLocalIdentifier>1234567890</ns:CriticalCareLocalIdentifier>
```
False Positive PCD leakage

- With a large number of records, short PCD fields are bound to produce false positive PCD leaks.
- Examples:
  - Date of birth
  - Short numeric identifiers (<6 characters)
- Some PCD fields contain just 1 character
False Positive PCD leak Detection

- Algorithm based on probability of PCD leak being a coincidence
- Probability is influenced by:
  - Type of PCD field (e.g. date of birth can have multiple formats)
  - Length of PCD and candidate fields (e.g. shorter PCD fields are more likely to be leaked accidentally)
  - Contents of PCD and candidate fields (whether it is only digits or alphanumeric characters)
  - Number of PCD leaks in in the interchange
  - Total number of records in the interchange
- Threshold value is used to determine whether leak is false positive
Binomial Distribution

- Discrete probability distribution of the number of successes in a sequence of $n$ independent experiments
- Defined by two parameters: number of experiments $n$, and probability of success $p$.
- Example:
  - We toss a fair coin 5 times.
    - Heads = success, tails = failure;
    - $n=5$
    - $p=0.5$
    - What is the probability that there were exactly 3 heads?
      - Mass function:
    - What is the probability that there were at most 3 heads?
      - Cumulative distribution function:

$$P(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n$$

$$F(x; p, n) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$$
Our problem can be modelled as a binomial distribution, where:

- "experiment" is finding out if the given PCD field is found in a given candidate field;
- "successful" outcome is a record where PCD field is included in the candidate field;
- "unsuccessful" outcome is a record where PCD field is not included in the candidate field;
- number of experiments is the total number of records;
- probability $p$ is the probability that a randomly generated PCD field includes a randomly generated candidate field.
Testing the algorithm

- Test interchanges were generated to support different test scenarios:
  - Large interchange with false positive date of birth leakage
    - All identifiers are randomly generated strings;
  - Large interchange with false positive spell number leakage
    - All identifiers are randomly generated strings; spell number is modified to contain 4-digit values
  - Large interchange with false positive patient name leakage
    - A small number of records are modified so that patient name is included in some candidate field
  - Interchange with obvious PCD leakage (PCD leak in every record)
  - Interchange with less obvious PCD leakage (a small number of true PCD leaks in a large interchange)
Example results

- 8000 records with random PCD and non-PCD fields generated
- Out of 8000 records in an interchange:
  - 17 date of birth fields with length 4 leaked in critical care local identifier with length 10
    - Likelihood of false positive: 99.94%
  - 8 spell number fields with length 12 leaked in unique identifier with length 27
    - Likelihood of false positive: almost 0%
  - 5 spell number fields with length 4 leaked in provider reference number with length 8
    - Likelihood of false positive: 37%
Threshold Effect

Interchange with 8000 records.
PCD field: date of birth
PCD field length: 4
Candidate field: provider reference number
Candidate field length: 8

<table>
<thead>
<tr>
<th>Probability Threshold</th>
<th>Maximum number of allowed false positive leaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05 (5%)</td>
<td>32</td>
</tr>
<tr>
<td>0.1 (10%)</td>
<td>30</td>
</tr>
<tr>
<td>0.2 (20%)</td>
<td>28</td>
</tr>
<tr>
<td>0.3 (30%)</td>
<td>26</td>
</tr>
<tr>
<td>0.5 (50%)</td>
<td>24</td>
</tr>
</tbody>
</table>
Threshold Effect

Interchange with 20000 records.
PCD field: spell number
PCD field length: 6
Candidate field: CDS unique id
Candidate field length: 27

<table>
<thead>
<tr>
<th>Probability Threshold</th>
<th>Maximum number of allowed false positive leaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001 (0.1%)</td>
<td>4</td>
</tr>
<tr>
<td>0.01 (1%)</td>
<td>3</td>
</tr>
<tr>
<td>0.05 (5%)</td>
<td>2</td>
</tr>
<tr>
<td>0.3 (30%)</td>
<td>1</td>
</tr>
<tr>
<td>0.5 (50%)</td>
<td>0</td>
</tr>
</tbody>
</table>
Redaction Overview

- Occurs late in the Pipeline
  - Testing shows that the point at which redaction occurs does not affect results.
  - This could be an issue if derivations were based on redacted fields.
- Data on disk is never updated/redacted - only the generated reports
- Low impact on resources (all the heavy processing done up-front)
- The redaction message is currently “REDACTED”, but can be anything
  - All candidate fields are currently strings, but code could handle date or integer redactions
- The threshold for handling false-positives can be modified at this stage - no need to re-run interchange files.
Redaction Example
Conclusion and next steps

- Longer the field, the more confident it is a true positives
- Threshold can be set low without making the algorithm too lenient for large interchanges
- Doesn’t matter when in the workflow redaction is applied
- Next steps
  - Evaluate interchanges with small number of records
  - Evaluate short fields and present results
  - Test algorithm with Live Data
  - Implement false positive logic without redaction
  - Decide what redaction replacement value to apply to each field
  - Decide on whether to replace whole field or just the redacted value