SCOPE Joint Action
Stakeholder Event

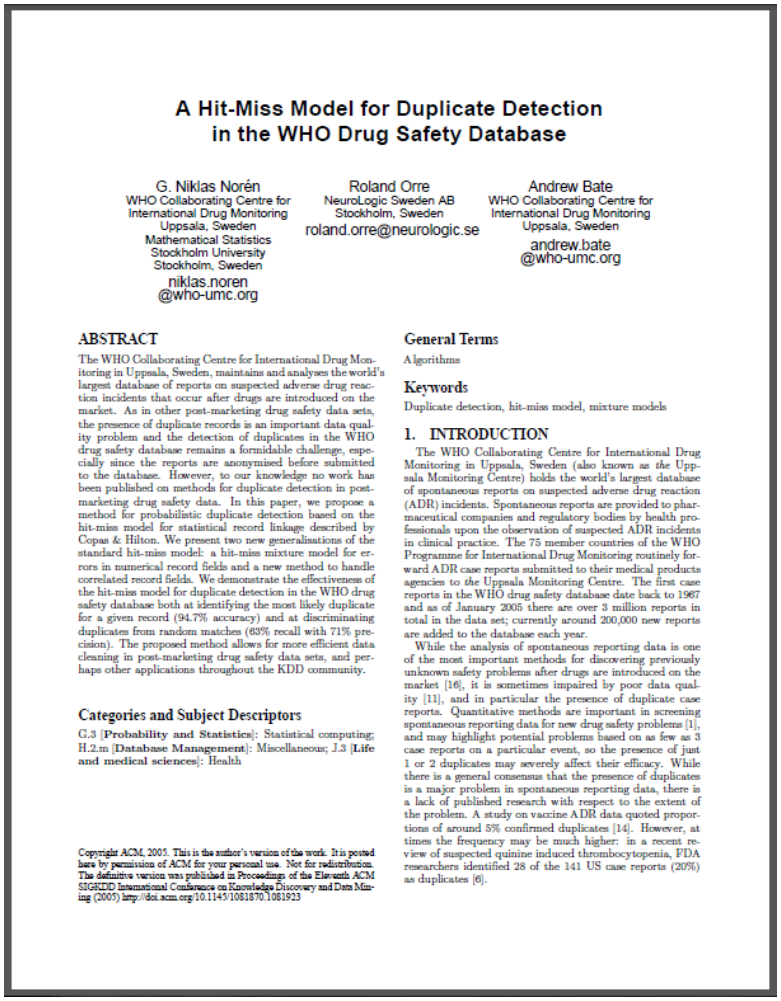
SCOPE Work Package 4
ADR Collection
Duplicate Detection

20 – 21 March 2017
London

Objectives

- Understand **potential approaches** to duplicate detection in a national or international database
- Describe the vigiMatch algorithm for duplicate detection and its effectiveness in duplicate detection

Approaches



A Hit-Miss Model for Duplicate Detection in the WHO Drug Safety Database

G. Niklas Norén
WHO Collaborating Centre for International Drug Monitoring
Uppsala, Sweden
Mathematical Statistics
Stockholm University
Stockholm, Sweden
niklas.noren@who-umc.org

Roland Orre
NeuroLogic Sweden AB
Stockholm, Sweden
roland.orre@neurologic.se

Andrew Bate
WHO Collaborating Centre for International Drug Monitoring
Uppsala, Sweden
andrew.bate@who-umc.org

ABSTRACT

The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden, maintains and analyses the world's largest database of reports on suspected adverse drug reaction incidents that occur after drugs are introduced on the market. As in other post-marketing drug safety data sets, the presence of duplicate records is an important data quality problem and the detection of duplicates in the WHO drug safety database remains a formidable challenge, especially since the reports are anonymised before submitted to the database. However, to our knowledge no work has been published on methods for duplicate detection in post-marketing drug safety data. In this paper, we propose a method for probabilistic duplicate detection based on the hit-miss model for statistical record linkage described by Copas & Hilton. We present two new generalisations of the standard hit-miss model: a hit-miss mixture model for errors in numerical record fields and a new method to handle correlated record fields. We demonstrate the effectiveness of the hit-miss model for duplicate detection in the WHO drug safety database both at identifying the most likely duplicate for a given record (94.7% accuracy) and at discriminating duplicates from random matches (63% recall with 71% precision). The proposed method allows for more efficient data cleaning in post-marketing drug safety data sets, and perhaps other applications throughout the KDD community.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical computing; H.2.m [**Database Management**]: Miscellaneous; J.3 [**Life and medical sciences**]: Health

General Terms

Algorithms

Keywords

Duplicate detection, hit-miss model, mixture models

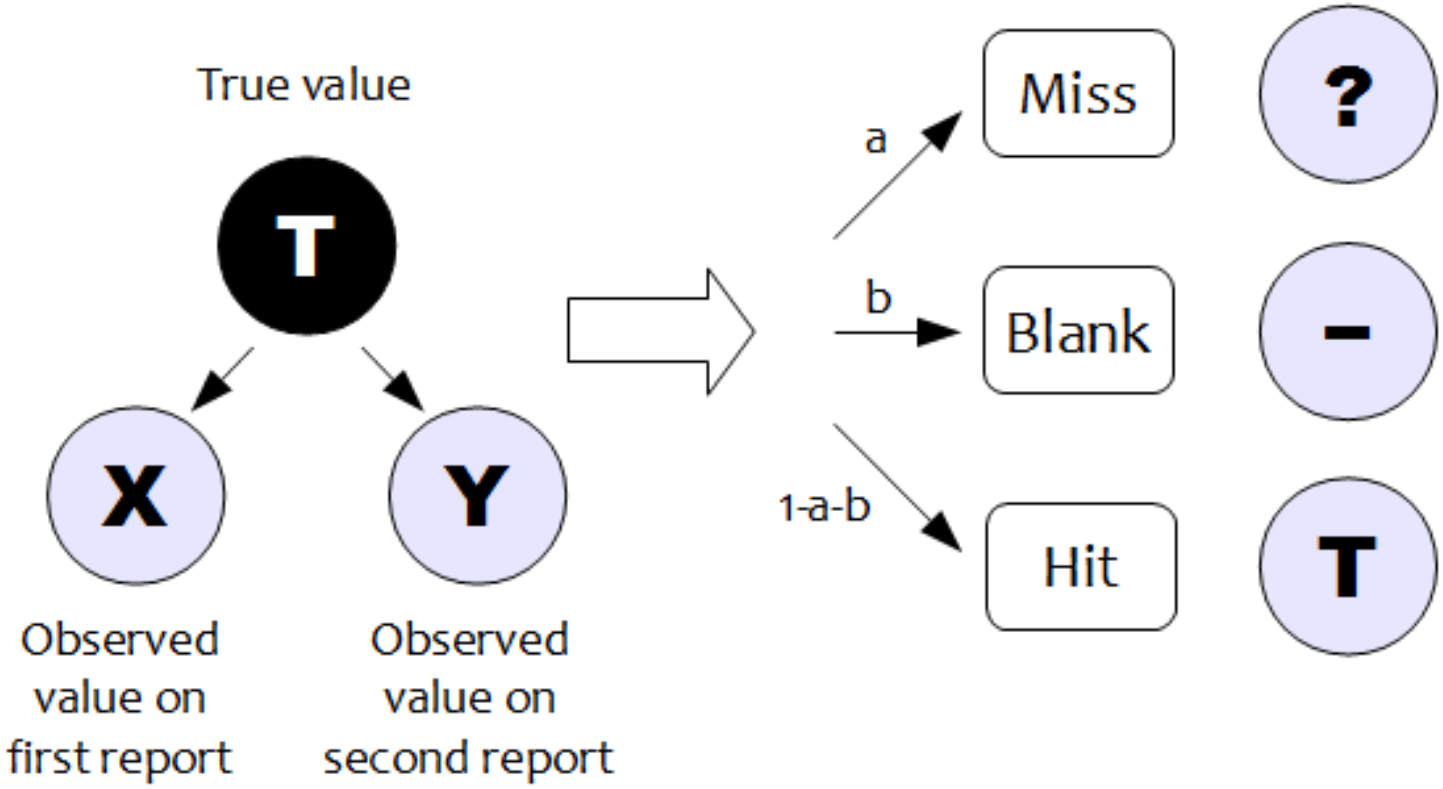
1. INTRODUCTION

The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden (also known as *the* Uppsala Monitoring Centre) holds the world's largest database of spontaneous reports on suspected adverse drug reaction (ADR) incidents. Spontaneous reports are provided to pharmaceutical companies and regulatory bodies by health professionals upon the observation of suspected ADR incidents in clinical practice. The 75 member countries of the WHO Programme for International Drug Monitoring routinely forward ADR case reports submitted to their medical products agencies to the Uppsala Monitoring Centre. The first case reports in the WHO drug safety database date back to 1967 and as of January 2005 there are over 3 million reports in total in the data set; currently around 200,000 new reports are added to the database each year.

While the analysis of spontaneous reporting data is one of the most important methods for discovering previously unknown safety problems after drugs are introduced on the market [16], it is sometimes impaired by poor data quality [11], and in particular the presence of duplicate case reports. Quantitative methods are important in screening spontaneous reporting data for new drug safety problems [1], and may highlight potential problems based on as few as 3 case reports on a particular event, so the presence of just 1 or 2 duplicates may severely affect their efficacy. While there is a general consensus that the presence of duplicates is a major problem in spontaneous reporting data, there is a lack of published research with respect to the extent of the problem. A study on vaccine ADR data quoted proportions of around 5% confirmed duplicates [14]. However, at times the frequency may be much higher: in a recent review of suspected quinine induced thrombocytopenia, FDA researchers identified 28 of the 141 US case reports (20%) as duplicates [6].

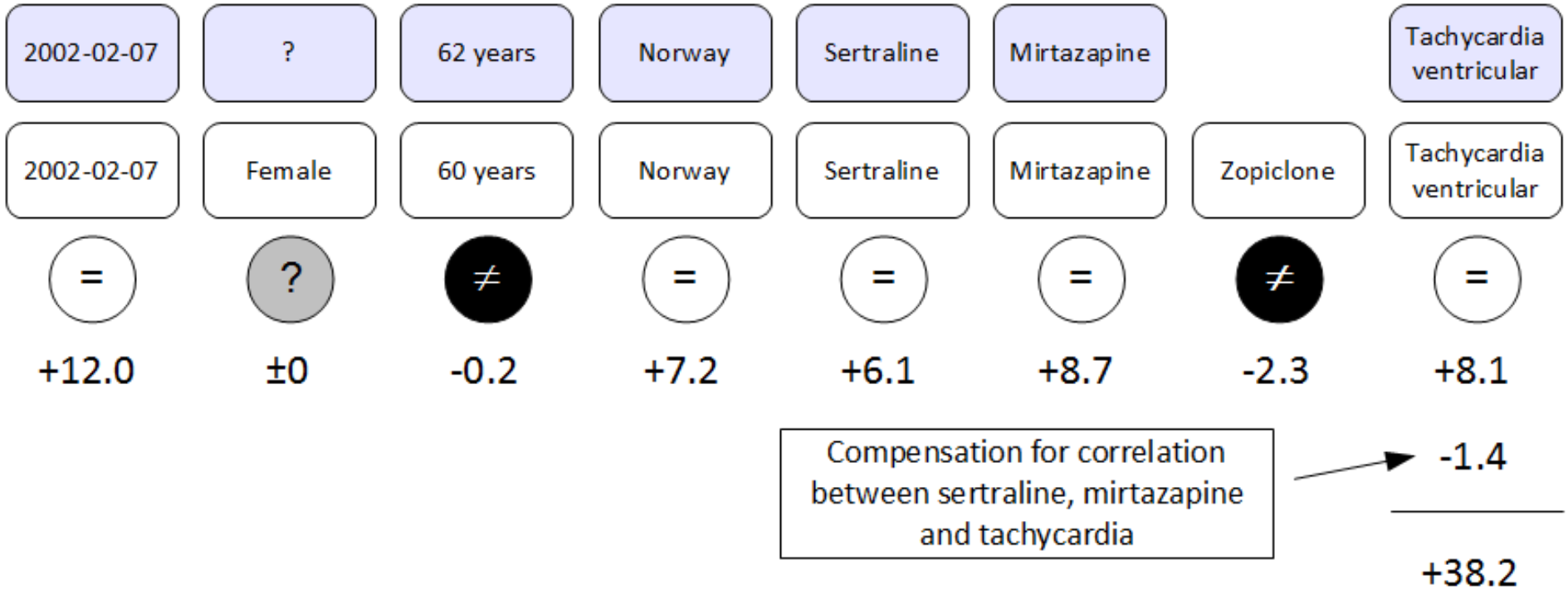
- Manual assessment
- Rule based approaches
- Score based approaches such as vigiMatch

The Hit-Miss Model



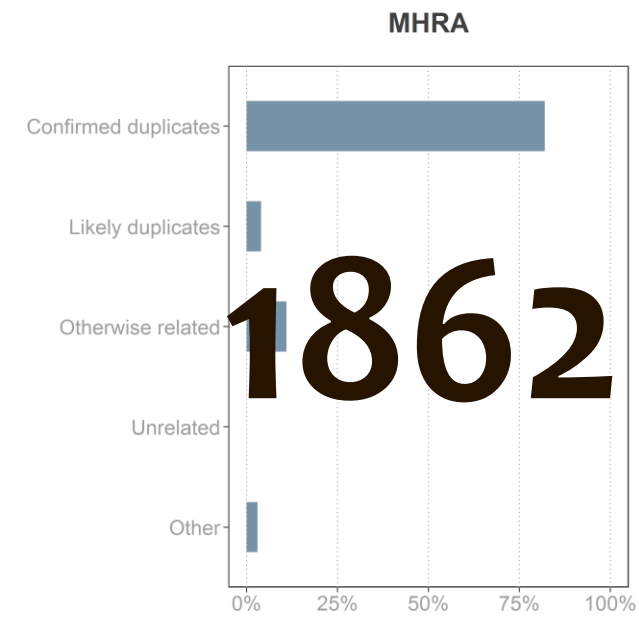
Norén et al.
Data Min Knowl Discov, 2007

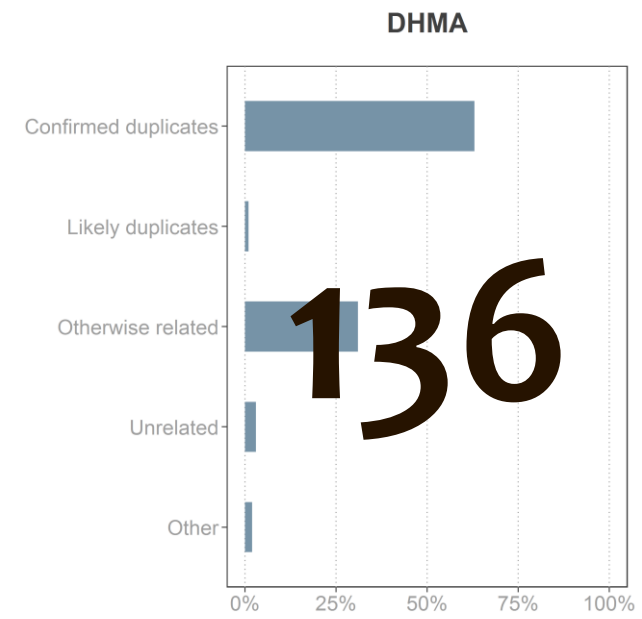
vigiMatch Scoring

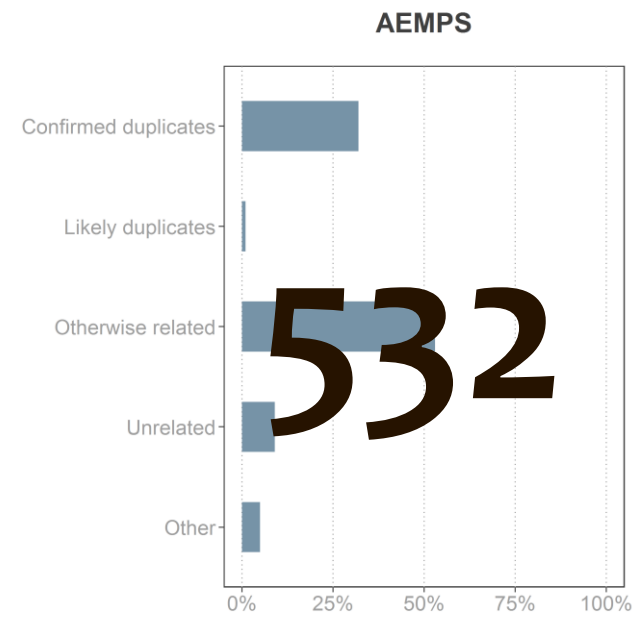


Norén et al.
Data Min Knowl Discov, 2007

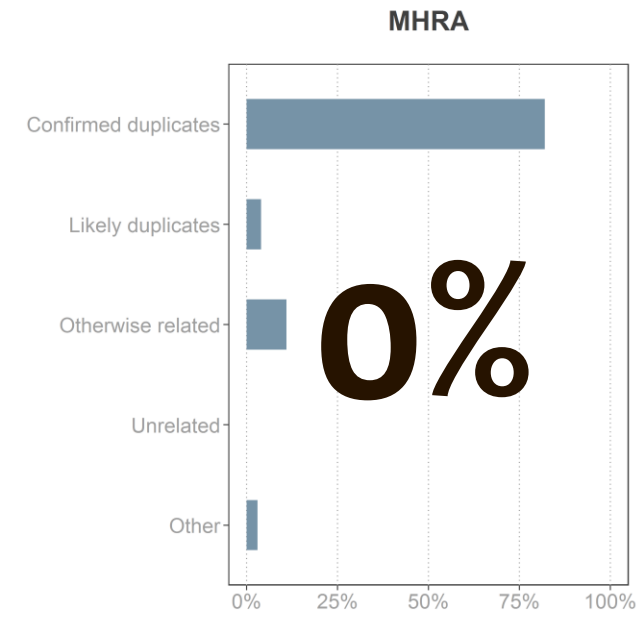
Relatively low rates of suspected duplicates

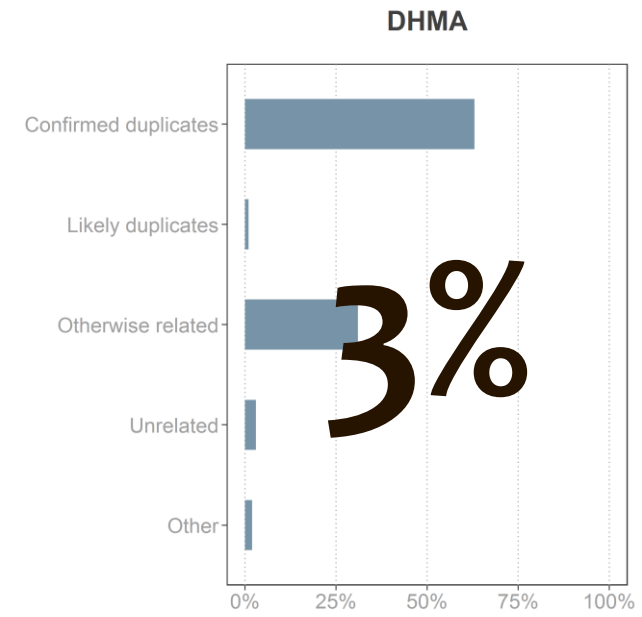


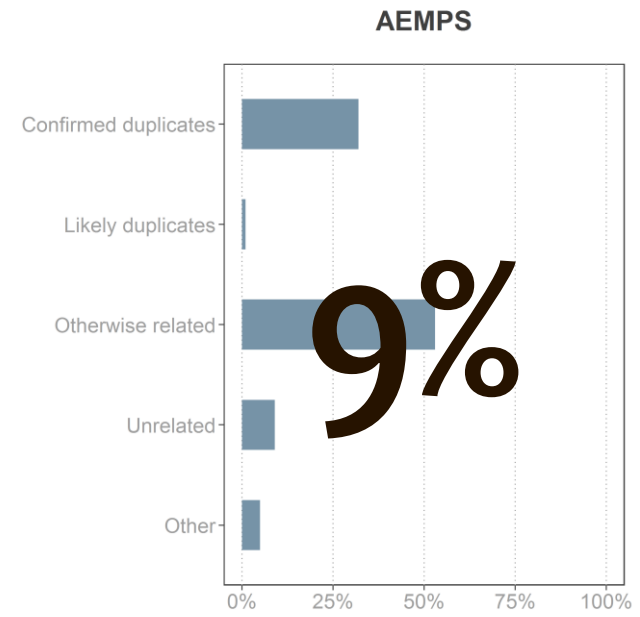




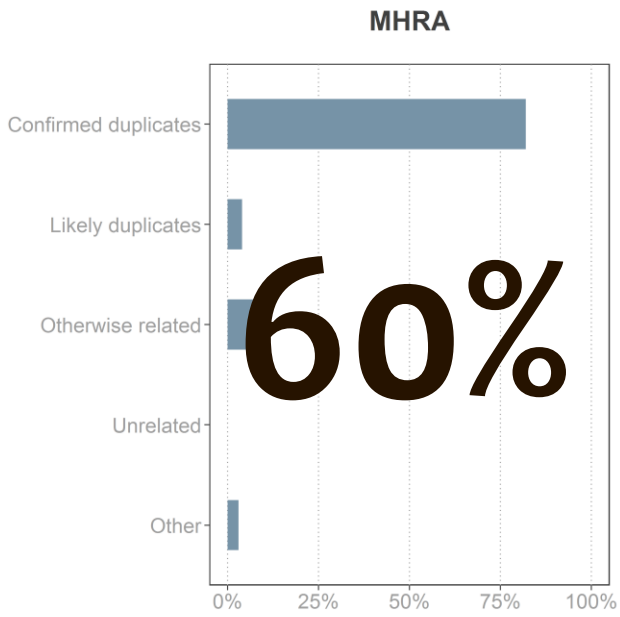
Low rates of false positives

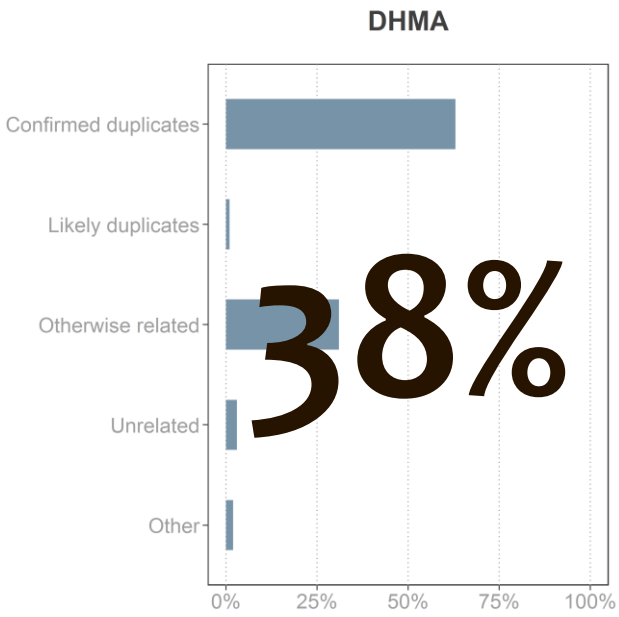


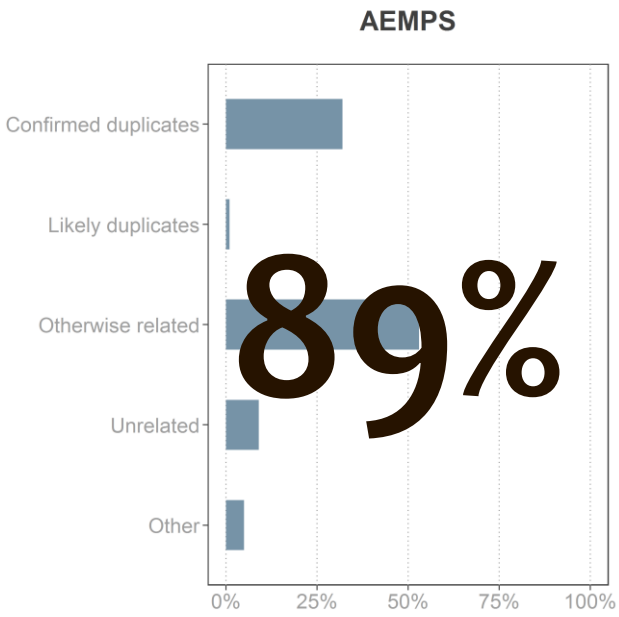




Most duplicates not previously known







Duplicate Detection Process

Take home messages

- There are a number of different approaches that you can take to duplicate detection dependent on your national setting
- vigiMatch is one potential approach validated as effective by PROTECT
- Complex algorithms do not necessarily need to be implemented in your own national database

Questions?

Contact:

scope@mhra.gsi.gov.uk