Uppsala Monitoring Centre (UMC) Box 1051, SE-751 40 Uppsala, Sweden +46 18 65 60 60, www.who-umc.org



Leveraging free text information to detect duplicates in COVID-19 vaccine adverse event reports

Erik Turesson (Uppsala University, Uppsala Monitoring Centre) and Jim W. Barrett (Uppsala Monitoring Centre)

Introduction

There are almost 35 million adverse event reports (AER) in VigiBase, the world's largest global database of AERs. These reports come from many sources, and sometimes duplicate reports are submitted for the same event. This negatively impacts both statistical and manual signal detection. Duplicates are often nonidentical, making them difficult to recognise automatically. Here we present a novel machine learning-based

Lessons from COVID-19

The COVID-19 vaccine rollout led to an unprecedented number of AERs being sent to VigiBase. The vaccinations happened over a short period of time, and in homogeneous populations (e.g., age groups were often vaccinated at the same time). This caused

tool for identifying duplicate COVID-19 vaccine AERs.

Report A

Age: **23**

Country: **US**

Sex: Male

Date: 23 Feb

Adverse Events: Weight Increase

Narrative: Cough following vaccination,

weight increased by 8.5 kg in 2 weeks

Report B

Age: **N/A**

Country: **UK**

Sex: Male

Date: 23 Feb

Adverse Events: Cough, Weight Increase

Narrative: Increase in weight

by <mark>8.5 kgs</mark> in 14 days

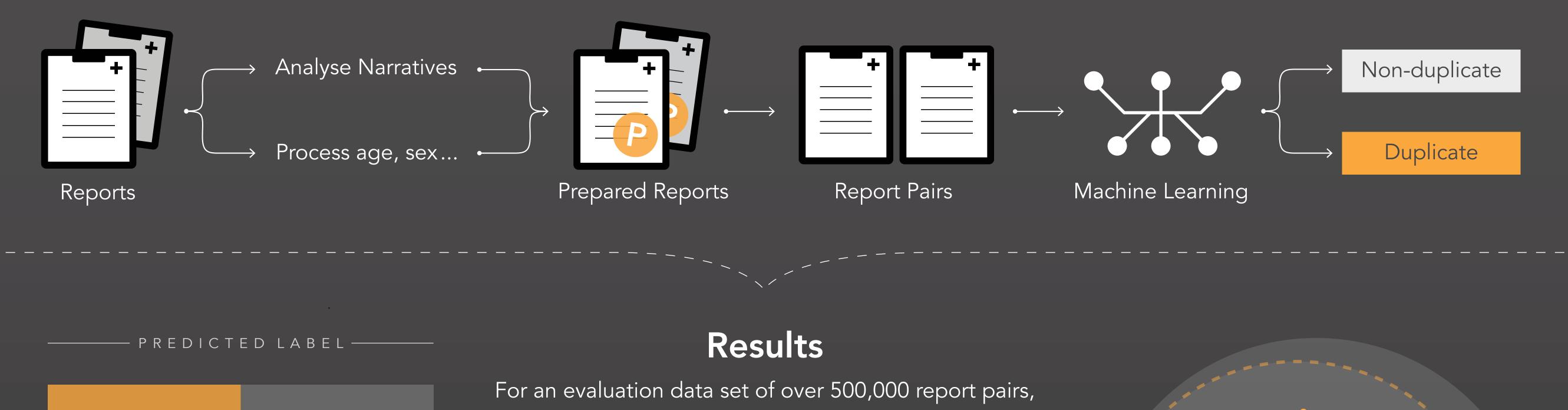
A fabricated example of duplicate reports. Note the discrepancies between them.

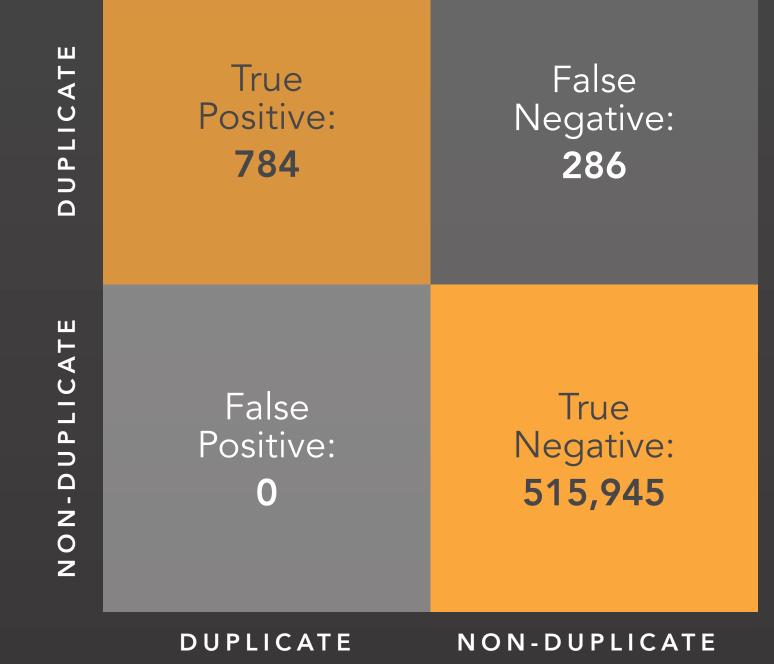
traditional deduplication methods to become ineffective. In order to effectively identify duplicates, we had to look to other parts of the AERs, including the narrative.

> Duplicates don't always look like twins.

Method

To assess whether a pair of reports are likely to be duplicates, we look at the similarity in, for example, age or the reported adverse events and pass them through a machine learning model trained on confirmed duplicate pairs. We also use state-of-the-art language models to measure the similarity between the two narratives. This process is summarised in the pipeline below.





В

Ш П our method found 73% of 1,070 true duplicate pairs. It also did not falsely identify any non-duplicates as being duplicates. The complete results are shown in the confusion matrix to the left.

Since our labelled data was subject to some bias, for an additional evaluation, we took all 11,756 COVID-19 vaccine AERs in VigiBase relating to hearing disorders. We applied our model to them and found 1,328 suspected duplicates and had 3 reviewers agree or disagree with the predictions for a subset of these pairs. The reviewers identified 87% of pairs predicted as duplicates as likely to be true duplicates. 37%

Humans agreed with 87% of the model's predictions