

Critical appraisal of artificial intelligence *for rare-event recognition: principles and pharmacovigilance case studies*

G. Niklas Norén  , Eva-Lisa Meldau , Johan Ellenius 

Uppsala Monitoring Centre, Uppsala, Sweden

niklas.noren@who-umc.org

Table 1. Intended use and test sets for use cases referred to in the paper

VigiBase pregnancy algorithm

- **Intended use:** Identifying adverse event reports related to pregnancy for individual case review or subsequent statistical analyses.
- **Scope:** All deduplicated VigiBase reports up to January 2023.
- **Out of scope:** The method does not analyze unstructured free text fields, which may contain relevant pregnancy information. Paternal exposures are not considered.
- **Reference standard:** Positive controls: reports with evidence suggesting they refer to cases where drug was administered to a pregnant person, before or during pregnancy/labour, with or without adverse event in the pregnant person or the foetus/prenatally exposed child. Negative controls: reports without such evidence, including reports where drugs were administered to a father, drug administered to the pregnant person postpartum or during lactation (unless administered also during pregnancy), drug administered to the child after birth, drug administered to non-pregnant person.
- **Prevalence of positive controls:** 3.2% of the reports in the downsampled data set of VigiBase were annotated as related to pregnancy. Based on pre-study results, the prevalence of pregnancy cases in VigiBase has been estimated to be around 1%.
- **Annotation process:** The annotations were performed by a team of experienced pharmacovigilance assessors following a predefined guideline. Inter-annotator agreement was quantified using Fleiss' kappa to be 83%, indicating a very good level of agreement.
- **Ambiguity:** Borderline cases, where pregnancy status was unclear or inconsistently documented, posed challenges for annotators, contributing to uncertainty in the reference standard. Ambiguity in the annotation process primarily stemmed from limitations in the structured data available in case

reports. In many instances, pregnancy-related information was present only in free-text narratives, which are not processed by the algorithm (and missing altogether from many reports due to restrictions in how these data is shared).

vigiMatch duplicate detection method

- **Intended use:** Highlighting suspected duplicate reports for subsequent human review and/or removal of suspected duplicates prior to statistical analyses of adverse event reporting data.
- **Scope:** VigiBase adverse event reports (27 million for medicines and 1.8 million for vaccines) available in January 2023.
- **Out of scope:** Nearly 5 million reports related to COVID-19 vaccines were excluded (due to their overrepresentation and atypical reporting patterns)
- **Reference standard:** Positive controls: Possible/likely duplicates; Negative controls: Non-duplicates (model-specific precision tests additionally included a second subcategory of negative controls corresponding to Otherwise related reports).
- **Prevalence of positive controls:** The prevalence of duplicate adverse event reports is not known and may vary. A recent study by the US FDA found 20% of the reports in the FAERS database to be suspected duplicates (33). Since the analysis of duplicates is at the level of report pairs, if 20% of the reports in a database of 40 million reports have a (single) duplicate, the prevalence of duplicates among all possible pairs of reports is expected to be 1 in 200 million ($0,2 \cdot 40 \cdot 10^6 / (40 \cdot 10^6)^2$).
- **Annotation process:** Recall study based on already annotated data from four different national regulatory authorities (for which annotation process is not well-described). Annotations for model-specific precision tests followed annotation guideline developed for study and included as supplementary material in Barrett et al (2025) (23). All pairs in the precision tests were doubly annotated to determine inter-annotator agreement.
- **Ambiguity:** Many reports pairs in VigiBase are not possible for human operators to classify with confidence as duplicates or non-duplicates because of the limited available information on reports. Inter-annotator agreement was estimated to a Cohen-Kappa score of 0.67, which suggests moderate to substantial agreement between human annotators, reflecting this ambiguity.

UMC redaction of narratives method

- **Intended use:** Highlighting possible person names for human review and redaction or automated redaction of possible person names in case narratives as a privacy-preserving measure.
- **Scope:** Yellow Card data, with case narratives written in English (UK) as they arrive at the pharmacovigilance organisation
- **Out of scope:** Excluding narratives derived from electronic health records as they were known to have lower expected prevalence of person names, and therefore be of less interest for redaction.
- **Reference standard:** Positive controls: NAME tokens; negative controls: NON-NAME tokens
- **Prevalence of positive controls:** 71 out 5,042 randomly sampled narratives (1.4%) included at least one person name, and in total 179 tokens were annotated as NAME compared to 263,272 as NON-NAME (0.07%).

- **Annotation process:** De novo annotations performed by four different annotators according to annotation guideline developed for purpose of study (see Meldau et al supplementary material). Lack of double annotations and of study of intra-annotator agreement acknowledged as a limitation.
- **Ambiguity:** There were difficult cases involving short tokens that could not be unambiguously classified (despite consultation with native English speakers from the UK, one of whom is a medical doctor), illustrating the fundamental difficulty and ambiguity of the task. When not possible to determine with confidence, tokens were annotated as NAME (aligned with false negatives being considered more costly than false positives).