Detecting Duplicates in Adverse Event Reports for *Cannabis sativa* and *Mitragyna speciosa* (Kratom) Products

Sanya B. Taneja, MS, Xiaotong Li, MS, Maryann R. Chapin, PharmD, Sandra L. Kane-Gill, PharmD, Richard D. Boyce, PhD
University of Pittsburgh, Pittsburgh, PA, USA

Introduction

There has been a recent sharp rise in the usage of products derived from the botanicals *Cannabis sativa* (cannabis) and *Mitragyna speciosa* (kratom). With concerns about potential adverse events, methods from pharmacovigilance, such as analyses of spontaneous safety reports submitted to the FDA Adverse Event Reporting System (FAERS), could be helpful for post-marketing safety surveillance. Unfortunately, duplicate adverse event reports are a common issue and very few methods exist for deduplication[1]. This issue gravely affects botanical pharmacovigilance due to an overall sparsity of reports compared to drugs in FAERS. Primary report identifiers are not reliable as reporting of the same events with different identifiers is common, especially for events that have also been reported in the literature[1]. Prior automated deduplication efforts include literature reference normalization[1] and, separately, probabilistic record matching (VigiMatch algorithm in WHO VigiBase)[2]. The goal of this study was to develop a method that extends VigiMatch for deduplication in FAERS and evaluate the methods with a focus on cannabis and kratom reports.

Methods

We developed an algorithm that starts with the probabilistic VigiMatch algorithm and performs additional deduplication using literature references, demographic characteristics, and drug and adverse drug reaction (ADR) information. The `lit_ref` field in FAERS (filled in for about 5% of FAERS reports) is a text field containing the author, title, and journal name of the associated publication available from September 2014 onwards. Where possible, these references were normalized to PubMed IDs using the PubMed citation matcher. Custom identifiers were assigned to non-PubMed indexed articles using string matching and cosine similarities between pairs of literature references in the entire database. All cannabis and kratom-related reports were queried from FAERS and deduplicated based on 1) identification by Vigimatch (in the Vigibase database), or 2) the literature reference, age, sex, weight, event date, reporter country, and the sum of numeric concept identifiers for drugs and ADRs that had been mapped to RxNorm and MedDRA respectively[3]. For evaluation, manual deduplication was conducted by two independent reviewers for all cannabis and kratom reports queried from FAERS.

Results

The FAERS database contained 13,966,995 reports (counting all report versions) from January 2004 to June 2022, including duplicate reports. Literature references were included in 4.7% of the reports and were normalized to PubMed IDs (68.4%) and custom identifiers (31.6%). Table 1 shows the deduplication results for cannabis and kratom reports. The extended algorithm identified 459 and 31 non-overlapping duplicates for cannabis and kratom respectively. While nearly doubling the capture of duplicate reports, manual evaluation showed that the method still missed a large proportion of duplicate reports (47.7% for cannabis and 37.5% for kratom).

Table 1. Deduplication results for cannabis and kratom.

	Duplicates: VigiMatch (%)	Additional duplicates from new algorithm (%)	Manually Identified Duplicates (%)	Total reports
Cannabis	207 (3.1)	252 (3.8)	886 (13.2)	6680
Kratom	18 (3.2)	13 (2.3)	50 (8.8)	570*

* Note: 168 kratom reports were not included in this analysis due to an unexpected change in the terms used for kratom in FAERS from Quarter 4 2021.

Discussion and Conclusions

This is the first study to focus on deduplication in FAERS for botanicals combining probabilistic matching, literature references, and adverse event reports details. While the literature reference was useful for cannabis and kratom reports, combining multiple approaches resulted in better deduplication. Unfortunately, manual review showed that extreme duplication remains an issue requiring further technical development. In future work, we will improve the method and report on its impact on pharmacovigilance signal investigation for all botanical products in FAERS using embeddings.

References

1. Hung E, Hauben M, Essex H, Zou C, Bright S. More extreme duplication in FDA Adverse Event Reporting System detected by literature reference normalization and fuzzy string matching. Pharmacoepidemiology and Drug Safety. 2023;32(3):387–91.
2. Tregunno PM, Fink DB, Fernandez-Fernandez C, Lázaro-Bengoa E, Norén GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug safety. 2014;37:249–58.
3. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data. 2016 Dec;3(1):160026.