Automating Adverse Event Extraction from EMA Summary of Product Characteristics

Robert ANCUCEANU[a,*], Doina DRĂGĂNESCU[a], Bogdan TAMBA[b]

[a] Carol Davila University of Medicine and Pharmacy, Faculty of Pharmacy, 6 Traian Vuia Street, 020955, Sector 2, Bucharest, Romania
[b] Advanced Research and Development Center for Experimental Medicine (CEMEX), "Grigore T. Popa" University of Medicine and Pharmacy, University Street No. 16, 700115 Iasi, Romania; Department of Pharmacology, Clinical Pharmacology and Algesiology, "Grigore T. Popa" University of Medicine and Pharmacy, University Street No. 16, 700115 Iasi, Romania
E-mail: robert.ancuceanu@umfcd.ro; doina.draganescu@umfcd.ro; bogdan.tamba@umfiasi.ro

* Author to whom correspondence should be addressed;

Abstract
Systematically extracting adverse event (AE) data from Summary of Product Characteristics (SmPC) documents is often necessary in exploring drug safety with computational means (e.g. for building QSAR models). The European Medicines Agency (EMA) makes available the SmPCs in a pdf format, which complicates this process. Inconsistent formatting (at least two different ways of structuring tables), multi-page tables, variable number of tables included in section 4.8, and varied terminology, combined with ambiguous frequency data, create significant barriers to reliable, large-scale (semi)automated analysis. We built a two-stage extraction workflow in R that integrates table-based and text-based methods. Using the "tabulapdf" package, we first retrieved AE tables spanning multiple pages, automatically removed repeated headers, and merged fragmented tables. We then applied a rule-driven text parser focused on 10 monoclonal-antibody-related AEs (anaphylaxis, cytokine release syndrome, angioedema, urticaria, rash, pyrexia, hypersensitivity, bronchospasm, hypotension, pruritus). The parser handles terminology variability through synonym mapping, prioritizes longer strings to avoid partial matches, and uses context-sensitive frequency detection at both line-level and section-level frequency indicators. Frequencies are mapped to a standardized 0-6 scale (0=absent, 1=unknown frequency, 2-6=very rare to very common). Outputs from the extraction tool are standardized, providing detailed metrics: frequency scores for each AE, aggregated counts of all AEs and serious cases (including anaphylaxis, cytokine release syndrome, and angioedema), and the peak frequency among reported AEs. The method is robust in processing nested tables, preserving the connection between frequencies and AEs despite line breaks, and correctly distinguishes between AEs that are absent from those with unreported frequencies. Performance validation challenges are addressed, alongside comparisons with available open-source and proprietary tools. The pipeline allows consistent and reproducible extraction of adverse event information from EMA SmPCs, supporting scalable comparative safety and pharmacovigilance analyses.

Keywords: Automatic Extraction; R language; SmPC; tabulapdf; QSAR models.