

Eureka – Finding a way to harvest the data in the medicinal product information

Luis Correia Pinheiro^{1,3}, Benedicte Cappelli², Jan Kors³, Xavier Kurz¹, Peter Arlett¹, Peter Rijnbeek³

1 Data Analytics and Methods Taskforce, European Medicines Agency, Amsterdam, The Netherlands

2 Pharmacovigilance Office, European Medicines Agency, Amsterdam, The Netherlands

3 Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

1. Objective

Build and validate a natural language processing (NLP) tool that extracts adverse drug reactions (ADR) from the English Summary of Product Characteristics (SmPC) of Centrally Authorised Products (CAPs) in the European Union (EU).

2. Methods

A symbolic NLP application titled EurEKA - EU's Product Information Entity extraction and Knowledge acquisition - was developed using R and Python. It downloads and parses SmPCs in PDF format and uses a bespoke dictionary based on the Medical Dictionary for Regulatory Activities (MedDRA) to identify ADRs. Entities are mapped to MedDRA. Results were compared to a database of manually curated ADRs published by the Pharmacoepidemiologic Research on Outcomes of Therapeutics by a European Consortium (PROTECT).

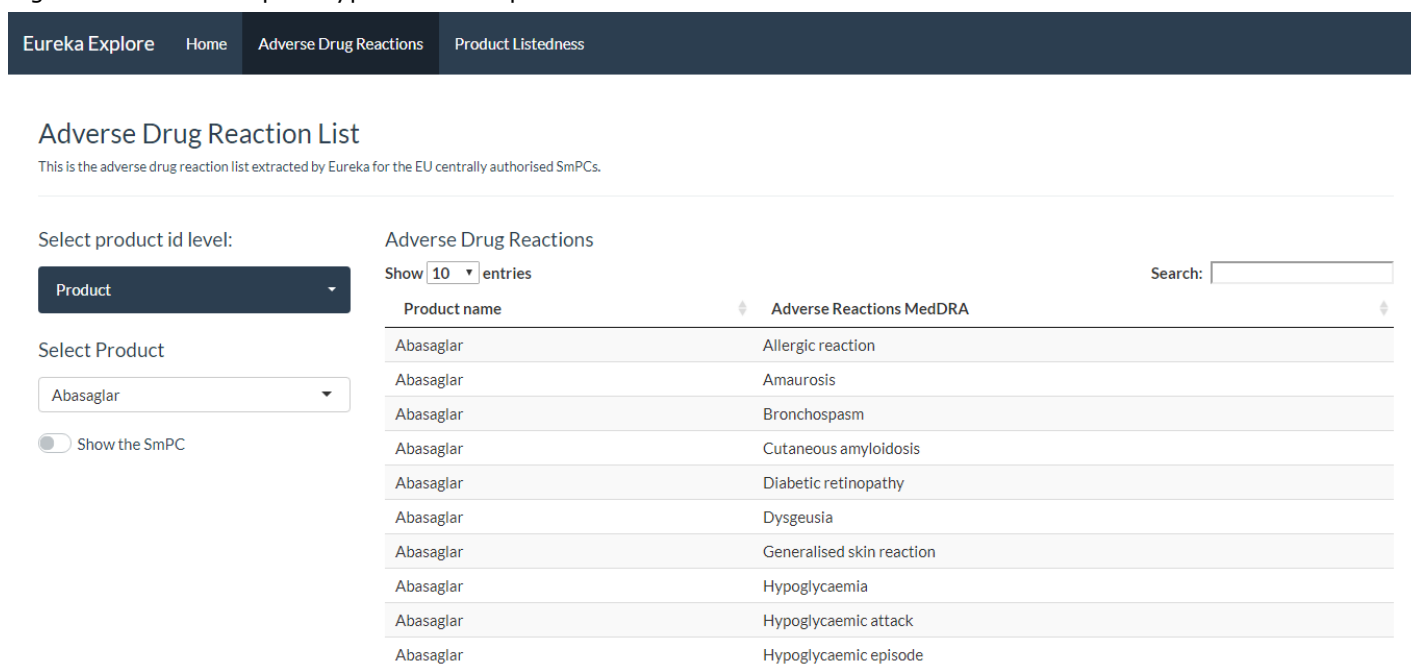
3. Results

A total of 1136 SmPCs were processed and performance was assessed on 910 of these. The macro-averaged performance was F1 0.81, Precision 0.80, and Recall 0.84. Micro-averaged performance was F1 0.84, Precision 0.82, Recall 0.85

4. Discussion

EurEKA achieved a performance equivalent to the best published performances. Analysis of a sample of SmPCs revealed missing terms in the PROTECT database suggesting performance is likely to be higher. The chosen method did not require annotating data and facilitates the implementation of multilingual methods by utilising available MedDRA translations. EurEKA was built with a human-in-the-loop module (EurEKA Extract) to correct extraction errors, and an analytical/query module (EurEKA Explore) (Fig. 1).

Fig 1. Screenshot of prototype Eureka Explore



The screenshot shows the 'Eureka Explore' web application. The navigation bar includes 'Eureka Explore', 'Home', 'Adverse Drug Reactions', and 'Product Listedness'. The main content area is titled 'Adverse Drug Reaction List' and includes a subtitle: 'This is the adverse drug reaction list extracted by Eureka for the EU centrally authorised SmPCs.' Below this, there are controls for 'Select product id level' (set to 'Product') and 'Select Product' (set to 'Abasaglar'). A 'Show the SmPC' toggle is present. The main table, titled 'Adverse Drug Reactions', shows 10 entries with columns for 'Product name' and 'Adverse Reactions MedDRA'. A search bar is located at the top right of the table area.

Product name	Adverse Reactions MedDRA
Abasaglar	Allergic reaction
Abasaglar	Amaurosis
Abasaglar	Bronchospasm
Abasaglar	Cutaneous amyloidosis
Abasaglar	Diabetic retinopathy
Abasaglar	Dysgeusia
Abasaglar	Generalised skin reaction
Abasaglar	Hypoglycaemia
Abasaglar	Hypoglycaemic attack
Abasaglar	Hypoglycaemic episode

3. Conclusions

EurEKA's performance is at the level of the best performing tools to extract ADR data and can be used to facilitate the extraction of ADR data.