# Not only pharmacovigilance data: the use of biochemical features in drug adverse events prediction

Pietro Belloni & Nicholas Tatonetti

Statistical methods and models for complex data — Poster Session: 21 September 2022

## Introduction

**Side effects of drugs** are a major cause of morbidity and mortality around the world. Therefore, careful monitoring of drug safety is essential to detect adverse drug events (ADEs) that may follow the administration of a drug. Many drugs' ADEs are discovered during clinical trial phases, particularly during phases II and III, but the relatively low sample size used in those stages causes a variety of infrequent effects to go unnoticed. Therefore, it is extremely important to identify associations between drugs and adverse events during the post-marketing phase (**phase IV**) is extremely important.

## State of the art: disproportionality models

Current methods used by pharmacovigilance institutes to predict ADEs are based on the statistical analysis of **spontaneous databases** (like FAERS) also known as disproportionality models. Adverse effect reporting is spontaneous: physicians, researchers, pharmacists - sometimes even patients - voluntarily report adverse events.

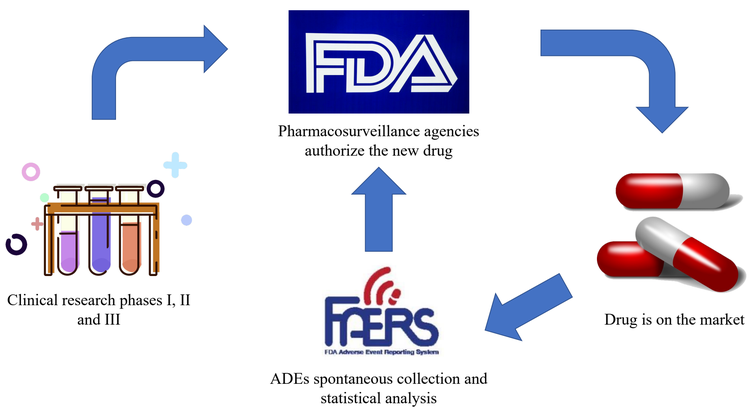


Figure 1. Use of spontaneous databases in pharmacovigilance.

Currently, pharmacovigilance agencies use a variety of **disproportionality models**, usually classified as Frequentist models (or Classical models), Bayesian models, Regression models and Machine Learning models. Some recent studies have compared the performance of different models on a gold standard, showing that Bayesian models, particularly those that introduce a shrinkage component, perform better. Nonetheless, using only spontaneous data, model performance remains moderate (AUC < 0.70).

| Model | Class | AUC |
|---|---|---|
| Bayesian Confidence Propagation NN | B | 0.69 |
| Gamma-Poisson Shrinkage | B | 0.68 |
| Reporting Odds Ratio | F | 0.66 |
| Proportional Reporting Ratio | F | 0.65 |
| Logistic Regression | R/M | 0.66 |
| Random Forest | R/M | 0.52 |

Table 1. Performance of some common disproportionality models on the OMOP Gold Standard Database. B: Bayesian, F: Frequentist, R/M: Regression/Machine Learning. Adapted from Pham et al., 2019.

## Limitations of current approach

What leads complex statistical models developed specifically for spontaneous data to still have moderate performance? By their nature, spontaneous data are **extremely biased**. The main problems are:

- **No control data**: since only cases are reported, there are no data from patients who have not taken drugs.
- **Under-representation**: only a small fraction of ADRs are reported.
- **Publicity bias**: a sudden increase in reports is due to a publication or news story related to the drug in question.
- **Weber effect**: after some time since the drug was introduced on the market, doctors get used to the side effects and stop reporting them.
- **Confounding variables**: often not reported in pharmacovigilance databases.
- **Lack of labeled data (gold standard)**: leads to difficulty in training machine learning models.

All of these biases lead disproportionality models to produce inaccurate estimates if they only use spontaneous data.

## Biochemical data: an alternative data source

The use of endogenous, unbiased data sources can increase the performance of ADEs prediction. To support classic spontaneous pharmacovigilance data, we chose to retrieve data from **biochemical structure** of drugs.

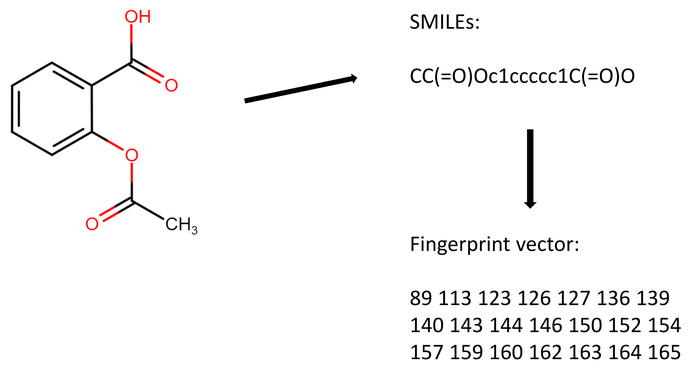


Figure 2. Both SMILES strings and Fingerprint vectors are derived from the drug structure using the CDK Java framework.

We extracted both SMILES strings and MACCS Fingerprint vectors from the active ingredient of the drugs. We focus here only on SMILES strings because they allow us to map a greater amount of information from the drug structure.

→ **SMILES** (simplified molecular input line-entry system) is a chemical notation specifically designed for computer use by chemists.

## SMILES as a chemical language

SMILES can be seen as a **language**, with a specific vocabulary and grammar rules, representing atoms, molecules, and bonds:

1. We used the strings as input for a BERT-like transformer model (ChemBERTa) capable of creating a large ($\sim$ 700) **embedding space**, where each chemical compound is mapped (Chithrananda et al., 2020).
2. The model masks 15% of each string and learns to predict by **auto-completion** masked atoms, groups of atoms and bonds.
3. **ChemBERTa**, avaiable on Hugging Face, has been pre-trained on 250k SMILES strings from the `zinc15` database.
4. The resulting embedding space represents a set of **latent variables** capable of describing a chemical compound.

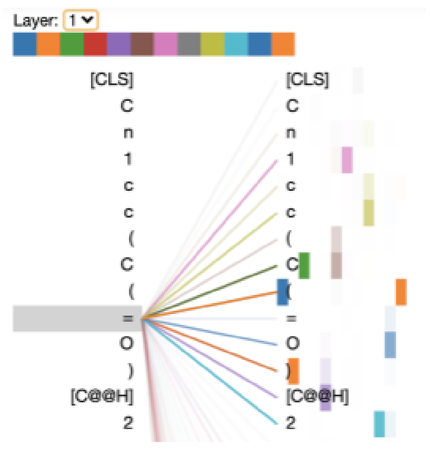


Figure 3. Graphical representation of a ChmBERTa attention layer.

## Results

We joined the set of features obtained from the embedding space with the **2019 FAERS data**. Then, we used the resulting dataset to predict the presence of ADE with a Support Vector Machine model. Finally, we evaluated the model using the area under the ROC curve and the area under the precision-recall curve. We also obtained confidence intervals for the evaluation metrics using bootstrap replications.

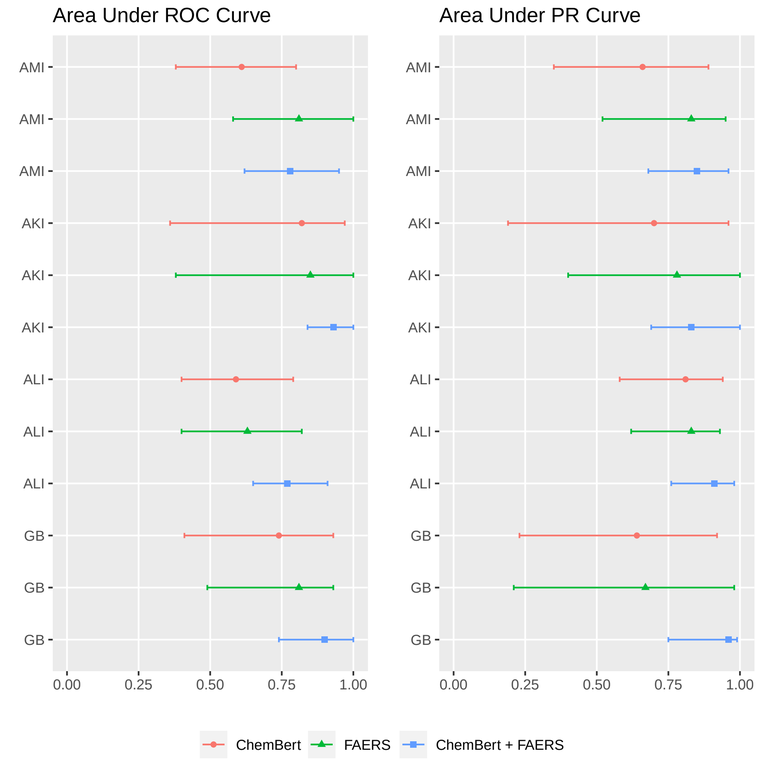


Figure 4. Result of the Support Vector Machine classifier on the OMOP Gold Standard Database. AMI: Acute Myocardial Infarction, AKI: Acure Kidney Injury, ALI: Acute Liver Injury, GB: Gastrointestinal Bleed.

In conclusion, using the joint set of features ChemBERTa + FAERS produces a higher **predictive power** than using only one of the two sets. Also, AUCs are higher than those previously shown.

## Conclusions

- Statistical analysis of spontaneous data alone can be used in the prediction of ADE, but there is **room for improvement**.
- Prediction power can be increased with data from **chemical structure of the drugs**.
- We interpret SMILES strings as a language and embed them in a **latent feature space** created by a BERT-like transformer model (ChemBERTa).
- This space of latent features has been joined to FAERS spontaneous data. This set of features allows a Support Vector Machine classifier to predict drug-ADE associations with **higher performance**.

## References

Pham, M., Cheng, F., Ramachandran, K. (2019). A Comparison Study of Algorithms to Detect Drug–Adverse Event Associations: Frequentist, Bayesian, and Machine-Learning Approaches. *Drug Safety*, **6**, 743 – 750.

Chithrananda, S., Grand, G., Ramsundar, B. (2020). ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint*.

## Contact information

**Pietro Belloni**, PhD Student
Department of Statistical Sciences, University of Padua
pietro.belloni.1@phd.unipd.it

**Nicholas Tatonetti**, Associate Professor
Department of Biomedical Informatics, Columbia University
npt2105@cumc.columbia.edu

@ www.tatonettilab.org