EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

AI in Pharmacoepidemiology

Opportunities for the European Regulatory Network

ENCePP Plenary, December 1st, 2023

Luis Pinheiro, EMA, Data Analytics and Methods Taskforce
With thanks to Valentijn de Jong, SNE, EMA, Data Analytics and Methods Taskforce

Content

- Brief introduction to AI
- Types of AI – Analytic vs Generative
- Promise of AI
- Use cases in Pharmacoepidemiology

Brief introduction to AI

AI is a catch-all term for a collection of novel digital technologies and methods

Allow machines to extract rules from data *(i.e., learn)*

Predict, classify or cluster similar patterns using those rules or execute autonomous actions

AI is data-driven

1. **Needs data** (data protection /security)
2. **Typically* needs *supervision*** (establish ground truth)
3. **Predicts/ classifies the past** (performance drops)
4. **Bias in data become rules** (ethical issues)
5. **Needs Human intelligence** (design issues)

Artificial intelligence | Types of AI

Analytic AI (Aim: group similar things or be accurate)		**Generative AI** (Aim: mimic)
Analyse, recognise patterns, predict, probabilistic classification, etc.	Capabilities	Generate content, structure text, support cognitive work. *Derives from Analytic AI.*
Automation, analysis, routine decision making, domain-specific interaction	General use	Support cognitive work, content generation, querying and structuring data
Usually structured data	Input	Usually unstructured data

AI in Medicines Regulation | The Promise

Greater efficiency /productivity

- Automate processes
- Address scalability issues
- Leverage personal assistants (chatbots)

Do more things, faster

Reduce error

- Facilitate access to information
- Reduce human cognitive load

Provide information at the fingertip

Expose/Prepare data

- Transform text data into structured data
- Reduce dimensionality of data
- Confounding adjustment
- Imputation of missing data

Structure and summarise information

Expand insights

- Probabilistic phenotyping
- Synthetic control arms
- Clinical prediction modelling
- Heterogeneity of treatment effects

Predict probability of events

Exposing data | Named Entity Recognition (NER)

Description

NER (and extraction), is a task performed to identify and extract entities.

Entities are objects from some class, such as a disorder, a medicinal product, a person's name, etc.

Pharmacoepidemiology use cases

NER generally an intermediate task – e.g. to expose or present data in a research-ready format

Phenotyping/identification of a patient population

Example

A pregnant 30-year-old woman, had constipation three days after starting iron supplements.

Entity Linking

Mentions are detected with the standard pipeline's mention detector.

A pregnant ENTITY 30-year-old woman ENTITY , had constipation ENTITY three days ENTITY after starting iron supplements ENTITY .

Exposing data | Generative AI (GenAI)

Description

Large language models, facilitate extraction and structuring of information, particularly that which is extensively available in the training data.

LLMs can be used for NER have limitations

Pharmacoepidemiology use cases

Structuring data to be ingested, e.g. in a protocol.

Coding

Screening through information

Example

json Copy code

```json
{
  "description": "Anaphylaxis is a severe allergic reaction that can be
  "symptoms": {
    "skin": ["hives", "itching", "redness"],
    "respiratory": ["shortness of breath", "wheezing", "coughing"],
    "gastrointestinal": ["nausea", "vomiting", "diarrhea"],
    "cardiovascular": ["weak pulse", "palpitations", "dizziness"],
    "other": ["anxiety", "confusion", "loss of consciousness"]
  }
}
```

sql Copy code

```sql
SELECT sex, COUNT(pat_id) AS patient_count
FROM person
GROUP BY sex;
```

Data insights | Confounding control

Description

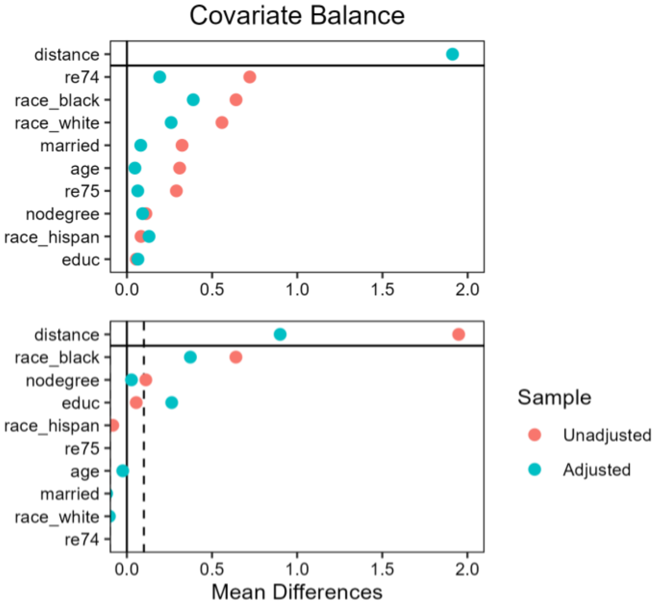
ML models may improve **propensity score estimation** through data-driven confounder selection, better modelling of non-linear effects and interactions

Overfitting less of a problem

Pharmacoepidemiology use cases

Structuring data that can be ingested, e.g. in a protocol or coding

Example



Data insights | Clinical prediction models

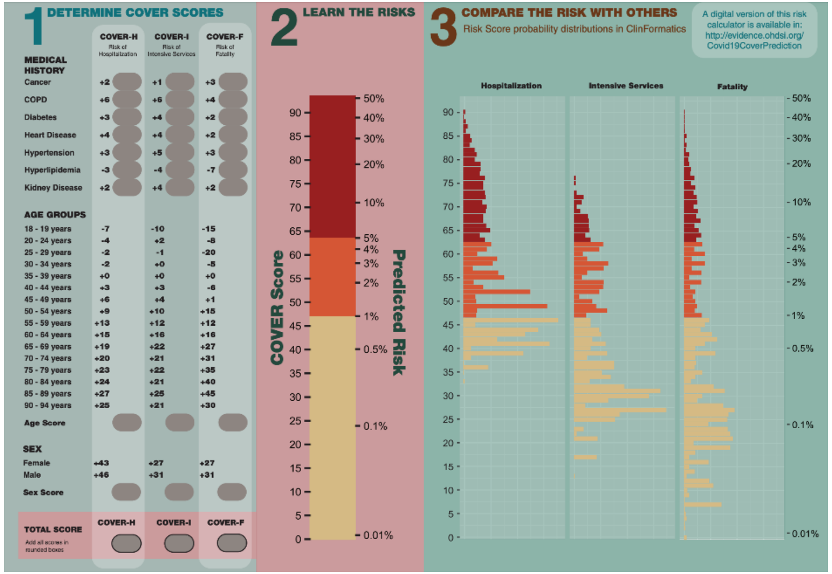
Description

Algorithmic risk scores not new to medicine and public health. ML models add the possibility to **predict probability of a clinical outcome** purely from a data-driven perspective

Pharmacoepidemiology use cases

Prediction of the probability of a clinical outcomes or behaviour (e.g. risk of abuse of opioids)

Example



Seek COVER: Development and validation of a personalised risk calculator for COVID-19 outcomes in an international network

Data insights | Probabilistic phenotyping

Description

In probabilistic phenotyping the ML identifies patterns in the characteristics of a population of interest and **predicts the likelihood that a patient is an element of that population**

Pharmacoepidemiology use cases

Cohort development – populations that share similar traits, particularly for phenotypes that have complex logic or are not directly identifiable from existing diagnostic/laboratory codes (e.g., gestational age)

Example

An Application of Machine Learning in Pharmacovigilance: Estimating Likely Patient Genotype From Phenotypical Manifestations of Fluoropyrimidine Toxicity

Luis Correia Pinheiro[1,*], Julie Durand[1] and Jean-Michel Dogné[2,3]

Dihydropyrimidine dehydrogenase (DPD)-deficient patients might only become aware of their genotype after exposure to dihydropyrimidines, if testing is performed. Case reports to pharmacovigilance databases might only contain phenotypical manifestations of DPD, without information on the genotype. This poses a difficulty in estimating the cases due to DPD. Auto machine learning models were developed to train patterns of phenotypical manifestations of toxicity, which were then used as a surrogate to estimate the number of cases of DPD-related toxicity. Results indicate that between 8,878 (7.0%) and 16,549 (13.1%) patients have a profile similar to DPD deficient status. Results of the analysis of variable importance match the known end-organ damage of DPD-related toxicity, however, accuracies in the range of 90% suggest presence of overfitting, thus, results need to be interpreted carefully. This study shows the potential for use of machine learning in the regulatory context but additional studies are required to better understand regulatory applicability.

Data insights | Causal inference

Description

Machine learning identifies relations and patterns within data, not focused on causation but correlation

Targeted Maximum Likelihood Estimation allows use of ML with few assumptions on data distribution

Pharmacoepidemiology use cases

Being explored for association studies with high-dimensional data with complex, non-linear relationships

Example

ORIGINAL ARTICLE

OPEN

Targeted Maximum Likelihood Estimation for Pharmacoepidemiologic Research

Menglan Pang,[a,b] Tibor Schuster,[a,b] Kristian B. Filion,[a,b,d] Maria Eberg,[a] and Robert W. Platt[b,c,e]

Background: Targeted maximum likelihood estimation has been proposed for estimating marginal causal effects, and is robust to misspecification of either the treatment or outcome model. However, due perhaps to its novelty, targeted maximum likelihood estimation has not been widely used in pharmacoepidemiology. The objective of this study was to demonstrate targeted maximum likelihood estimation in a pharmacoepidemiological study with a high-dimensional covariate space, to incorporate the use of high-dimensional propensity scores into this method

Results: Through a real example, we demonstrated the double robustness of targeted maximum likelihood estimation. We showed that results with this method and inverse probability weighting differed when a large number of covariates were included in the treatment model.

Conclusions: Targeted maximum likelihood can be used in high-dimensional covariate settings. In high-dimensional covariate settings, differences in results between targeted maximum likelihood and inverse probability weighted estimation are likely due to sensitiv-

Summary

- Several, and increasing number of, opportunities to leverage AI for pharmacoepidemiology
  - Generative AI will likely increase productivity/efficiency but will likely still play a limited role in the actual analytics real-world data where data tends to be structured
  - Machine learning will become increasingly part of the analytical arsenal of the pharmacoepidemiologist but not likely to replace existing tools (e.g., for the most part, logistic regression models still are the best option)
- Experimentation and collaboration will be essential – a European Medicines Regulatory Network workplan including experimentation and fora for interactions with, e.g. Academics, is being prepared

Any questions?

Further information

luis.pinheiro@ema.europa.eu

Official address Domenico Scarlattilaan 6 • 1083 HS Amsterdam • The Netherlands
Telephone +31 (0)88 781 6000
Send us a question Go to www.ema.europa.eu/contact

Follow us on **@EMA_News**