SMAD-CC : SMArt Data for improved machine learning in Cancer Care

Paul MINCHELLA[1,2], Dounya BOURHANI[1,2], Eliel KATCHE[1], Stephane CHRETIEN[2], Guillaume METZLER[2], Loic VERLINGUE[1]

CRCL - Centre Léon Berard, 29 Rue Gabriel Sarrazin, 69008 LYON
ERICLAB Lyon2, 5 Avenue Pierre Mendès France, 69500 BRON

SHAPE-Med Lyon — Structuring one Health Approach for Personalized Medicine

FRANCE 2030

anr

CONTEXTE

SMAD-CC proposes to combine clinical, biological and molecular data with medical reports, for patients with various cancers. Our main objective is to improve the quality and completeness of data available to predict patient survival and evaluate the effects of medical interventions. This involves using various types of data at different points in the patient's history. We will then train our model on clinical reports using Natural Language Processing (NLP), and compared the performance of the data enrichment process. The Expected results include a demonstration of the added value of recent advances in robust statistics and machine learning theory. SMAD-CC is at the forefront of multimodal artificial intelligence for personalized cancer care, with many possible impacts on research and clinical application in this field.

INITIAL

```
Patient seen during hospitalization for treatment.

The clinical examination is unchanged.
The biological assessment is good.

Risk benefit:
We are therefore administering today on the VVP the first cycle of CABAZITAXEL at a
dose of 20 mg/m² for his adenocarcinoma of metastatic castration-resistant prostate.
I had already given him all the prescriptions.
I plan to have a CT scan and a bone scan after 4 treatment courses.
```

Clinical Data

PATIENT INFORMATION

- Cancer type : Prostate adk
- Concurrent treatments : Aspirin, Clopidogrel, Bisoprolol
- Symptoms : Pollakiuria, Hematuria
- Biological exams : PSA dosage, NFS
- Current performance status : PS 1
- Medical history :
  - diagnosis of a prostate tumor radical prostatectomy in may 2018
  - Biological recurrence in december 2018,
  - Radiotherapy of the prostate compartment in mars 2019 associated with hormonal therapy,
  - Re-increase in PSA under hormonal therapy with discovery of bony locations in september 2019 for which he had several radiotherapy sessions.
  - Hormone resistance in february 2020
  - ENZALUTAMIDE then chemotherapy with DOCETAXEL. Abiraterone Acetate, from february 2020.

Pharmacological Data

DRUGS DETECTION
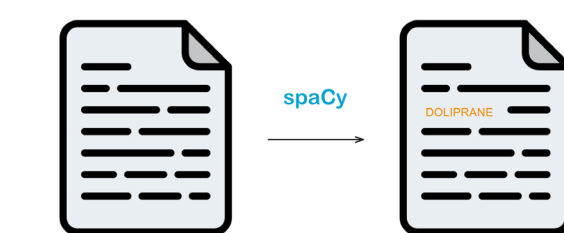
spaCy

Models	Precision	Recall	F_score
spaCy EDS-NLP	0.95	0.93	0.93
DeBERTa	0.44	0.56	0.49
BERT	0.11	0.27	0.15
Camembert	0.39	0.77	0.51
DrBert	0.35	0.75	0.48

TRANSFORMATION INTO SMILES

1. Construction of our dictionnary

ChEMBL + DRUGBANK =

Id	Smiles	Synonyms
1	CC(=O)NC1CCC(O)CC1	['ACETAMINOPHEN', 'PARACETAMOL', ...]
2	OC1CC(O)CC(O)C1	['PHLOROGLUCINOL', 'SPASFON', ...]
...	...	...
96206	C#CC1CCCC(NC2NCN...	['ERLOTINIB', 'TARCEVA']

2.a. Change the drugs with their SMILES

Dictionary → **93%** drugs are found

2.b. Use of LLM on the drugs that were not found in the dictionnary

Input: "Give me the components of LAMALINE" → LLM → Output:
- Paracetamol
- Opium
- Caffeine

Dictionnary
- CC(=O)NC1=CC=C(C=C1)O
- [H][C@@]12OC3=C(O)C=...
- CN1C=NC2=C1C(=O)N(C)C(=O)N2C

Molecular analysis reports

DATA EXTRACTION

Extraction of Molecular data with Artificial Intelligence assisted tools

DATA ENGINEERING

Patient	Sample	Chr	Strand	Pos	Ref	Alt
1	20243	12	+	25398285	C	T
1	20243	17	+	7578212	G	A
...	...	...	...	...	...	...
2	39992	7	+	140453136	A	T

VARIANTS ANALYSIS

OncoKB — ClinVar — CGI-Clinics

PATHOGENIC ALTERATIONS

Gene	Protein Change	Oncogenicity
AR	L702H	Pathogenic
TP53	R213*	Pathogenic
TMPRSS2-ERG	Translocation	Pathogenic

ENRICHED

```
Patient seen during hospitalization for treatment.
Karnofsky = 90%, PS = 1, Current weight = 77.0 kg, Body mass index = 25.73 kg/m²
Cancer type = Prostate adk, Symptoms = Pollakiuria, Hematuria, Biological exams = PSA dosage, NFS
Concurrent treatments = CC(=O)OC1=CC=CC=C1C(=O)O, COC(=O)[C@H](C1=CC=CC=[ ... ], CC(C)NCC(COC1=[ ... ]
Medical history = [ ... ]

The clinical examination is unchanged.The biological assessment is good.

Molecular analysis showed the presence of the following pathogenic alterations AR L702H, TP53 R213*
and TMPRSS2-ERG Translocation

Risk benefit:
We are therefore administering today on the VVP the first cycle of [H][C@]12[C@H](OC(=O)C3=CC=CC=C3)[ ... ]
at a dose of 20 mg/m² for his castration-resistant metastatic prostate adenocarcinoma

I had already given him all the prescriptions.
I plan to have a CT scan and a bone scan after 4 treatment courses.
```

Lyon 1 — Université Lumière Lyon 2 — HCL HOSPICES CIVILS DE LYON — Inserm — CENTRE LEON BERARD — VetAgro Sup — LE VINATIER PSYCHIATRIE UNIVERSITAIRE LYON METROPOLE — CNRS — Inria — INRAE — CPE LYON — International Agency for Research on Cancer / World Health Organization