Annotation Guidelines for Italian NSCLC Clinical Reports

25 Clinical Entities – Version 1.0 (November 2023)

Reference paper:
Domenico Paolo et al., 'Named Entity Recognition in Italian Lung Cancer Clinical Reports using Transformers', IEEE Access, 2023
DOI: 10.1109/BIBM58861.2023.10385778

1. General Annotation Rules

- Task: sequence labeling using BIO scheme
  - B–XXX = Beginning of entity XXX
  - I–XXX = Inside entity XXX
  - O = Outside (no entity)
- Entities can span multiple tokens
- Nested entities are forbidden
- Annotate the longest reasonable span
- Do NOT annotate entities extractable via rules: age, sex, smoking py, PD-L1, labs

2. The 25 Clinical Entity Types

Acronym	Entity	Description & Examples
CAN	Cancer	Primary tumor or metastasis → adenocarcinoma, carcinoma squamoso
COM	Comorbidity	BPCO, diabete, ipertensione
STA	Cancer stage	stadio IV, malattia metastatica
FAN	Focal anomaly	nodulo spiculato, lesione sospetta
POS	Anatomical position	lobo superiore destro, segmento 6
TPY	Therapy	chemioterapia, immunoterapia
DRU	Drug	cisplatino, pemetrexed, osimertinib

DOS	Dosage	500 mg, 75 Gy
DUR	Therapy duration	6 cicli, 12 mesi
TPL	Therapy line	seconda linea, terzo ciclo
TUP	Tumor progression	progressione, risposta completa
QHA	Quantity habits	20 sig/die, 30 py
EXA	Exam	TC torace, PET-TC
FRE	Frequency	ogni 21 giorni
PEV	Patient event	sospensione per tossicità
PSY	Symptom	dispnea, tosse
MAS	Mass	massa di 4 cm
MOR	Morphology	margini irregolari, aspetto spiculato
DAT	Date	15/03/2022
HIS	Histology	squamoso, adenosquamoso
FAM	Familiarity	madre con carcinoma
TNM	TNM	T2N3M1c
NRS	Pain scale	NRS 7/10
WEI	Weight	68 kg
HEI	Height	172 cm

3. Inter-Annotator Agreement

Token-level average: 0.98 ± 0.04
Entity-level average: 0.97 ± 0.08

4. Annotation Tool

Doccano was used for manual annotation.